High Resolution Single-Molecule Enzyme Dynamics Using Nanopores

Jonathan M. Craig

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jens H. Gundlach, Chair

Paul Wiggins

Anton Andreev

Program Authorized to Offer Degree:
Department of Physics

University of Washington

Abstract

High Resolution Single-Molecule Enzyme Dynamics Using Nanopores

Jonathan M. Craig

Chair of the Supervisory Committee:
Professor Jens H. Gundlach
Department of Physics

DNA is a molecule that contains the genetic information of all living organisms. DNA provides the instructions that the cell uses to construct the proteins which carry out the complex functions required for life to thrive. Enzymes are proteins that use chemical potentials to catalyze energetically unfavorable chemical reactions to perform various chemical and mechanical tasks ranging from muscle contraction to DNA packaging. In this thesis I focus on a class of enzymes called 'motor enzymes' which use the energy provided by ATP hydrolysis to move along a molecular track, such as DNA or RNA, and perform mechanical tasks such as unwinding double-stranded nucleic acids or building double stranded nucleic acids from single-stranded nucleic acids. Classically, enzyme were probed through biochemical methods that monitor a large number of reactions simultaneously. These methods are limited because enzymes operate near thermal energies, leading to asynchronous progression of the chemical reaction and therefore can only provide the average rate of the enzyme process, obscuring the finer details of enzyme activity. In the past 30 years, methods to monitor the reactions of single enzyme molecules have provided numerous insights into the function of motor enzymes, but these techniques lack the resolution to provide full details of how these molecules transduce chemical energy into mechanical work. In this thesis I present the development of Single-molecule Picometer Resolution Nanopore Tweezers (SPRNT), a single-molecule method developed from nanopore DNA sequencing for monitoring the movement of single

enzyme molecules on DNA at unprecedented spatiotemporal resolution using the biological nanopore MspA.

In SPRNT, a single MspA protein pore (termed a 'nanopore') in a phospholipid bilayer forms the only electrical connection between two salt solutions. A voltage applied across the membrane causes an ion current to flow through the nanopore. Negatively charged single-stranded DNA complexed to a motor enzyme is attracted into the nanopore by the electric field. The DNA passes through the pore until the motor enzyme, which is too large to fit through the pore, comes to rest on the rim of MspA. The DNA bases in the pore reduce the ion current flowing through the pore depending on the bases therein. The motor enzyme then moves along the DNA, causing DNA to move through the pore, leading to a series of stochastic ion-current amplitudes which simultaneously provide measurements of the kinetics of the enzyme and the DNA sequence. This method leads to a higher spatiotemporal resolution than any other single-molecule technique.

In this thesis I present my role in the development of SPRNT. In chapter 1 I introduce the relevant biomolecules and techniques used to examine them. In chapter 2 I discuss the development of SPRNT and quantify its spatiotemporal resolution. In chapters 3 and 4 I present the first enzyme dynamics studies done with SPRNT on the helicase Hel308, and use information from quantities that could not be measured previously to elucidate the precise details of Hel308 motion on DNA, and to determine the mechanism by which the DNA bases in Hel308 regulate its translocation on ssDNA. Chapter 5 contains concluding remarks and a discussion of the future of SPRNT.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Glossary vi

Chapter 1: Introduction 1
1.1 DNA, Proteins and Enzymes 1
1.2 Nanopore DNA Sequencing 4
1.3 Enzyme Kinetics 8
1.4 Single-Molecule Enzyme Kinetics: Imaging One Enzyme at a Time 12

Chapter 2: Sub-Angstrom Single-Molecule Measurements of Motor Proteins Using the MspA Nanopore 15
2.1 Introduction 15
2.2 Results 16
2.3 Discussion 28
2.4 Spatiotemporal Resolution of SPRNT 29

Chapter 3: Revealing the Kinetic Mechanism of a SF2 Helicase Using SPRNT . . 33
3.1 Introduction 33
3.2 Kinetic Methods 36
3.3 Results 38
3.4 Discussion 52

Chapter 4: Investigating the Effects of DNA Sequence on Hel308 Translocation on ssDNA 54
4.1 Introduction 54

4.2 Results . . . 56
4.3 Discussion . . . 63

Chapter 5: Conclusions . . . 65

Appendix A: Supplementary Information for chapter 2 . . . 80
A.1 Materials and Methods . . . 80
A.2 Experiment Statistics and DNA strands . . . 83
A.3 Elongation of DNA . . . 86
A.4 Automatic Consensus Generation from a Reference . . . 88
A.5 Hel308 Step Size Measurements . . . 91
A.6 ATP Titration of Hel308 . . . 93
A.7 Proposed 2-Step Mechanism and Hel308 Translocase Experiment . . . 95

Appendix B: Supplementary Information for chapter 3 . . . 98
B.1 Sources of Ion Current Modulation in Nanopore Experiments . . . 98
B.2 Conversion of Ion Current to DNA Position . . . 100
B.3 The Master Equation . . . 105
B.4 Statistical Analysis of Michaelis-Menten Parameters for f|f [ATP]-dependent Steps . . . 108
B.5 Analysis of Probability of Backwards Steps . . . 109
B.6 Comparing f|f and f|b [ATP]-dependent steps . . . 113
B.7 Derivation of a Dwell time Distribution Function for General f|b Steps . . . 114
B.8 Analysis of f|f and f|b [ATP]-independent Dwell time Distributions Using the AIC . . . 116
B.9 Voltage and Temperature Variation . . . 120
B.10 Calculation of Average Dwell Time of f|f [ATP]-dependent Steps Using the Steady-state Approximation . . . 122
B.11 Derivation of the Probability of a b|f Step for [ATP]-dependent Steps in Model 1 and Model 2 . . . 125
B.12 Estimation of Kinetic Parameters for [ATP]-dependent Steps . . . 131

Appendix C: Supplementary Information for chapter 4 . . . 138
C.1 Materials and Methods . . . 138
C.2 Calculation of p-value using 2-sample KS-Test . . . 138

LIST OF FIGURES

Figure Number		Page
1.1	DNA	3
1.2	Concept of Nanopore DNA Sequencing	6
1.3	MspA nanopore	7
1.4	Reproducibility of Nanopore Sequencing Reads	9
1.5	Basic Chemical Reaction Diagram	9
1.6	Reaction Network Diagram	10
1.7	Michaelis-Menten Mechanism	11
1.8	Techniques for Observing Single Enzymes	13
2.1	Schematic of SPRNT	17
2.2	DNA Position Measurements with SPRNT	18
2.3	SPRNT Analysis of Hel308 DNA Helicase	22
2.4	Sequence B Consensus	25
2.5	Sequence C Consensus	27
2.6	Resolving Steps in Noisy Data	30
2.7	Comparing Spatiotemporal Resolution of Single-molecule Techniques	32
3.1	SPRNT on Hel308 Helicase	35
3.2	Methods for Analyzing Enzyme Kinetics with SPRNT	37
3.3	Analyzing ATP and ADP Dependence of Hel308 Forwards Steps	39
3.4	Analyzing ATP and ADP dependence of Hel308 Backwards Steps	40
3.5	Probability of b|f [ATP]-independent Steps vs. DNA Position	41
3.6	Comparing f|f and f|b [ATP]-dependent Steps	42
3.7	Comparing f|f and f|b [ATP]-independent Dwell Time Distributions	43
3.8	Comparing f|f and b|f [ATP]-independent Dwell Time Distributions	45
3.9	Effects of Varying Voltage and Temperature on Hel308 Kinetics	47
3.10	Kinetic Model of Hel308 Translocation on ssDNA	50

4.1 Analysis of Hel308 dwell-time distributions for two DNA Strands 55
4.2 Experimental Scheme to Investigate Hel308 Sequence Dependence 57
4.3 Hel308 Translocation on Homopolymer Sequences 60
4.4 Determining sequence offset distance 62
4.5 Transporting backsteps between sequences and mutating the DNA sequence 64

A.1 DNA stretching in response to applied force 87
A.2 Result of Consensus Algorithm 89
A.3 Measuring Hel308 Step Size 92
A.4 ATP Titration of Hel308 94
A.5 Proposed Mechanism of 2-step Hel308 Translocase 96
A.6 Hel308 Translocase Experiment 97

B.1 Analysis of Ion Currents 100
B.2 Conversion of ion current to DNA position 103
B.3 A simple kinetic model to illustrate the master equation. 105
B.4 ATP Dependence of All f|f [ATP]-dependent Steps 110
B.5 Analysis of Michaelis Parameters 111
B.6 ADP dependence of All f|f [ATP]-dependent Steps 112
B.7 A kinetic model to analyze f|b steps. 115
B.8 All f|f [ATP]-independent Dwell Time Distributions 119
B.9 The kinetic model from main text figure 6a, reduced to a form that allows easy application of the steady-state approximation. 124
B.10 Calculating the Probability of b|f [ATP]-dependent Steps as a function of [ATP] and [ADP] 125
B.11 Fixed Ratio [ATP]:[ADP] = 1:4 Experiment 129
B.12 Comparing Kinetic Models by Analyzing Probability of b|f Steps 130
B.13 Calculating Kinetic Parameters 134
B.14 Alternative Model 1 135

LIST OF TABLES

Table Number Page

2.1 Comparing Single-molecule Techniques 31

A.1 Experiment Statistics For chapter 2 84
A.2 List of DNA Sequences Used in chapter 2 85

B.1 Experimental conditions and number of Hel308 events 104
B.2 Number of Measurements of f|f [ATP]-dependent steps 109
B.3 Using the AIC to Analyze f|f and f|b [ATP]-independent Steps 118
B.4 Best fit value of β to equation B.23 for the curves shown in figure 3.9. 121
B.5 List of Michaelis-Menten Parameters 136
B.6 Calculated Parameters For Model 1 137

C.1 List of DNA sequences used in chapter 4 139

GLOSSARY

DNA (DEOXYRIBOSE NUCLEIC ACID): The genetic material of living organisms. DNA consists of an alternating sugar-phosphate backbone, attached to which is one of four nucleotides: A,C,G,T. The order of bases determines many cellular processes.

ABASIC: DNA sugar phosphate group with no attached base.

RNA (RIBONUCLEIC ACID): Similar to DNA, with three of the same bases (A,C,G), and a fourth base (U) which is T without the methyl group. Each sugar contains an extra $2'$ hydroxyl group.

LIPID BILAYER: Two layers of lipid molecules which are organized with the hydrophobic tails oriented towards the center of the bilayer, and hydrophilic head groups oriented towards the exterior.

AMINO ACIDS: A polymer whose backbone is constructed from peptide bonds of amino groups to carboxyl groups. Attached to each part of the chain is an 'R-group' which is typically one of 20 different chemical structures, with varying charge, size, and solubility in water.

PROTEINS: Molecules constructed from chains of amino acids which perform cellular tasks. The order of the amino acids determines how the protein folds into more organized three dimensional structures.

MEMBRANE PROTEIN: Proteins which sit in the cell membrane, allowing for transport of materials into and out of the cell.

MSPA (MYCOBACTERIUM SMEGMATIS PORIN A): A goblet shaped membrane protein, whose geometry is useful for nanopore DNA sequencing.

ENZYME: A class of proteins which catalyze energetically unfavorable chemical reactions by converting small molecules without depletion of the enzyme.

MOTOR ENZYME: A class of enzymes which typically use ATP hydrolysis to generate motion along molecular tracks such as DNA or RNA.

TRANSLOCASE: Motor enzymes which move along DNA.

DNA POLYMERASE: Motor enzymes which construct dsDNA from a ssDNA template.

RNA POLYMERASE: Motor enzymes which locally unwinding dsDNA, and copy a template DNA strand into messenger RNA.

HELICASE: Motor enzymes which catalyze the unwinding of double stranded DNA for the purposes of replication.

ACKNOWLEDGMENTS

There are far too many people to whom I owe thanks for making it to this point. My PI Jens Gundlach taught me the importance of truly understanding the raw data, challenging my own conclusions, and how to effectively communicate my results. Andrew Laszlo and Ian Derrinton are inspirational mentors, fantastic scientists, and I am thankful for their guidance. Catherine Provost provided guidance and support throughout my graduate school career. Henry Brinkerhoff, Ian Nova, and Matthew Noakes are great colleagues and friends, and I owe them thanks for contributing to the work done in this thesis. Benjamin Tickman, Kenji Doering and Noah de Leeuw collected a significant amount of the data displayed here, and were a joy to work with. Brian Ross, Jenny Mae Samson, Kyle Langford, Hugh Higinbotham, Josh Bartlett, Samuel Klebanoff, Mark Svet, Katherine Baker, Jonathan Mount, Jasmine Bowman, and Sinduja Marx are others who I enjoyed working with. Our collaboration with Illumina on Hel308 enabled large fraction of the work done in this thesis. My first-year cohort in the physics department are great friends who helped me improve as a scientist and person from day 1. My family has been especially important in helping me through difficult times, especially the strength of my mother Evelyn, and my grandmother (Oma) Nelly. My brothers Aaron and Kevin have been incredibly supportive over this time as well. My uncle Roy Hammerling has been an inspiration to me since I was young, and I am thankful for having the opportunity to converse with him throughout my life. Also, thanks to Dwayne "The Rock" Johnson for keeping me stoked throughout my graduate program.

This work was supported through the National Institutes of Health, National Genome Research Institute (NHGRI) $ 1000 Genome Program Grant number R01HG005115.

DEDICATION

For Robert H. Craig.

Chapter 1

INTRODUCTION

Life is constructed from fundamental building blocks that are present in all living Terran organisms. DNA contains the genetic information of organisms, differentiating species and determining more specific character traits of individual members of a given species. In cells DNA is transcribed into RNA, which is translated into the proteins that perform the tasks by which life functions. The goal of this chapter is to introduce DNA, proteins and enzymes, and to introduce the experimental tools that I will use to examine these molecules at unprecedented resolution.

1.1 DNA, Proteins and Enzymes

DNA is a polymer chain whose backbone is made up of alternating sugar and phosphate groups. Attached to each sugar is one of four 'bases': adenine (A), cytosine (C), guanine (G), or thymine (T). A base together with the associated sugar-phosphate group is termed a 'nucleotide'. The sequence of bases is the genetic code that contains the instructions that govern cellular processes, making it of vital importance to determine the order of the bases. Each strand of DNA has an orientation, with the $5'$ end referring to the terminal phosphate group, and the $3'$ referring to the terminal hydroxyl group (figure 1.1). In cells, DNA consists of two anti-parallel strands which are attached by base-paired hydrogen bonds. A pairs with T by two hydrogen bonds, and C pairs with G by three hydrogen bonds. When DNA is a single polymer chain we refer to it as single-stranded DNA (ssDNA); when it is base-paired we refer to it as double-stranded DNA (dsDNA). ssDNA has a width of 1.2 nm and an internucleotide spacing of about 0.5 nm, while dsDNA has a width of 2.4 nm and an internucleotide spacing of 0.34 nm [1]. Importantly for the applications of this thesis, at

physiological conditions (pH $\approx$ 7.5) the phosphate group on each DNA nucleotide contains one negative charge. Thus, ssDNA is a charged molecule with a linear charge density of 1 e^- per nucleotide. Therefore, DNA is a molecule which can be manipulated by the application of electric fields.

RNA is a similar molecule to DNA, with three of the same bases (the fourth base, uracil is just thymine missing a methyl group), but the sugar group contains one additional hydroxyl group on the 2′ carbon. RNA has many functions in the cell from gene regulation [2] to acting as a molecular catalyst [3, 4]. RNA has been hypothesized to form the genetic basis of the first life forms on earth, due to its ability to act as both gentic information and as a molecular catalyst which can perform replication processes [5]. For the purposes of cellular replication, messenger RNA (mRNA) is a single-stranded RNA molecule copied from dsDNA. The messenger RNA is the code which the ribosome, a cellular factory complex, interprets to build proteins.

Proteins, like DNA and RNA, are polymer chains with a repeating backbone of peptide bonded carboxyl and amino groups, and attached side groups. The individual subunits of proteins are termed 'amino acids'. The ribosome determines which amino acid to add to a growing polymer chain by reading consecutive 3-nucleotide segments of mRNA. Proteins are constructed primarily from 20 different amino acids, whose side chains have various properties such as physical dimension, charge, and solubility in water. The order of amino acids determines how proteins fold into organized three-dimensional structures. Less water-soluble 'hydrophobic' amino acids tend to pack towards the center of the protein, while more water-soluble 'hydrophilic' amino acids tend to reside on the surface of the protein [6]. Many proteins fold into atomistically reproducible structures, enabling precise and reproducible biological functions. Many three-dimensional protein structures have been determined using experimental techniques such as X-ray crystallography [7] (figure 1.3) and Nuclear Magnetic Resonance (NMR, [8]).

An important class of proteins are called enzymes. Enzymes are protein catalysts that use the chemical energy gained by converting small molecules to catalyze energetically unfavor-

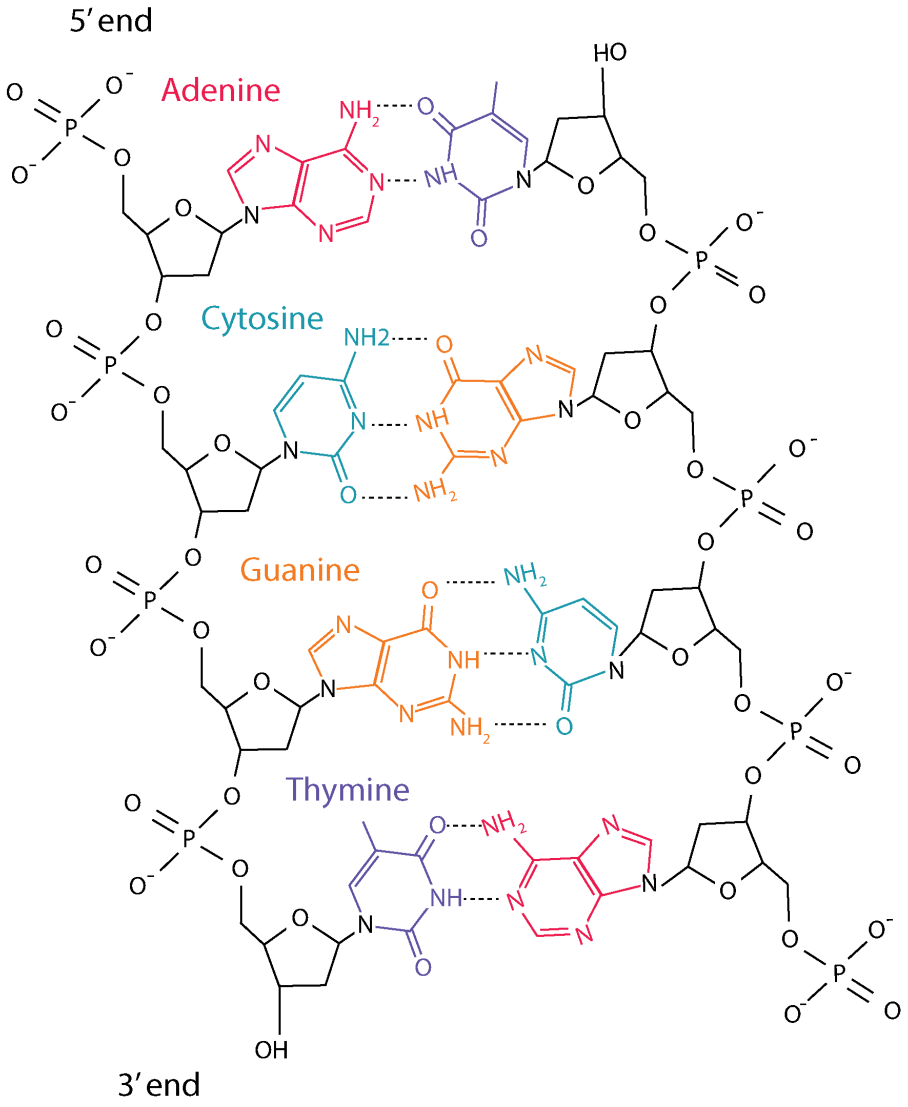

Figure 1.1: DNA

Chemical structure of double-stranded DNA. The bases are highlighted in color: adenine (red), cytosine (cyan), guanine (orange) thymine (purple). Hydrogen bonds between the two strands are indicated by black dashed lines.

able chemical processes such as the unwinding of dsDNA for cellular replication, the production of ATP, and other processes which would otherwise occur too slowly to be biologically useful. Enzymes are typically labeled by the suffix '-ase' (e.g. glycosylase, ATP-synthase).

For the purpose of this thesis, I will focus on a class of enzymes termed 'motor enzymes'. Motor enzymes use the energy gained from ATP hydrolysis (or similar processes) to produce directed motion along a molecular track and perform mechanical tasks. For example, DNA polymerases are enzymes which build double-stranded DNA from a single-stranded DNA template and individual nucleotide triphosphates (NTPs) [9]. RNA polymerases move along a double-stranded DNA template, locally unwind the dsDNA, and construct the mRNA that is used for building proteins [10]. DNA Helicases unwind double-stranded DNA so that DNA can be replicated[11]. The roles that motor enzymes play in cellular replication are of great importance to understanding molecular biology, and can have an impact in general health care. Defects in molecular motors have been implicated across a wide range of diseases such as hearing loss, polycystic kidney disease, and neurodegenerative diseases [12], while some viruses encode for their own motor-enzymes that can disrupt cellular function, and may be potential drug targets [13, 14].

1.2 Nanopore DNA Sequencing

The human genome contains three billion DNA base-pairs, which made determining its sequence one of the most impressive scientific achievements of the past century [15, 16]. Until recently it has been prohibitively expensive to sequence an individual's genome for clinical applications. However, next-generation sequencing technologies are bringing down the cost of DNA sequencing and increasing the rate at which DNA bases are sequenced, enabling personalized medicine [17, 18, 19]. I discuss one such technology here, nanopore DNA sequencing, which forms the experimental foundation upon which the rest of this thesis is built.

In nanopore DNA sequencing, a nanometer-scale opening in a membrane ('nanopore') forms the only electrical connection between two electrolyte solutions, termed *cis* and *trans*

[20]. A voltage applied across the nanopore causes an ion current to flow through the pore. Because DNA is a negatively charged molecule in solution, it is electrically drawn through the pore. The DNA moving through the pore will modulate the ion current flowing through the pore depending on the nucleotides in the pore. The DNA sequence could then be determined by simply measuring the ion current (figure 1.2).

Several different approaches have been taken to realize nanopore DNA sequencing experimentally. In one method a hole is formed in a solid-state material, such as silicon nitride or graphene, to act as a nanopore [21]. Another method uses biomolecules, with an artificial cell membrane ('lipid bilayer') as the membrane, and a membrane protein, such as α-Hemolysin ('αH') as the nanopore [22, 23]. The former method suffers from difficulty in reliably reproducing nanopores of a consistent size, and the nanopores tend to change in time [24, 25]. The University of Washington nanopore lab has pioneered the use of *Mycobacterium smegmatis porin A* ('MspA') as a biological nanopore for DNA sequencing [26, 27].

MspA (figure 1.3) is a goblet-shaped membrane protein with eight-fold symmetry, whose narrowest section ('the constriction') measures 1.2 nm wide, and 0.6 nm tall [28]. These physical dimensions are remarkably consistent with the spacing between single-stranded DNA bases, making MspA a probe that is highly sensitive to the DNA bases in its constriction. However, it was found that when DNA was introduced to the *cis* compartment, the DNA translocated through MspA too quickly to resolve any single-nucleotide sequence information [26]. While it was determined through clever experimental techniques that MspA had base-discrimination capability [27, 29], it was clear that a robust method for slowing down DNA translocation was required to make nanopore DNA sequencing practical.

The solution was provided by using the molecular-motor enzyme phi29 DNA polymerase ('phi29 DNAP') in conjunction with αH and MspA nanopores to control translocation of DNA through the nanopore [30, 31]. The result was that phi29 DNAP was able to control the DNA through the nanopore in stochastic, discrete, single-nucleotide steps that could be identified with the DNA sequence, enabling enzyme-controlled nanopore DNA sequencing at single-nucleotide resolution [31]. Enzyme-controlled nanopore sequencing reads are highly

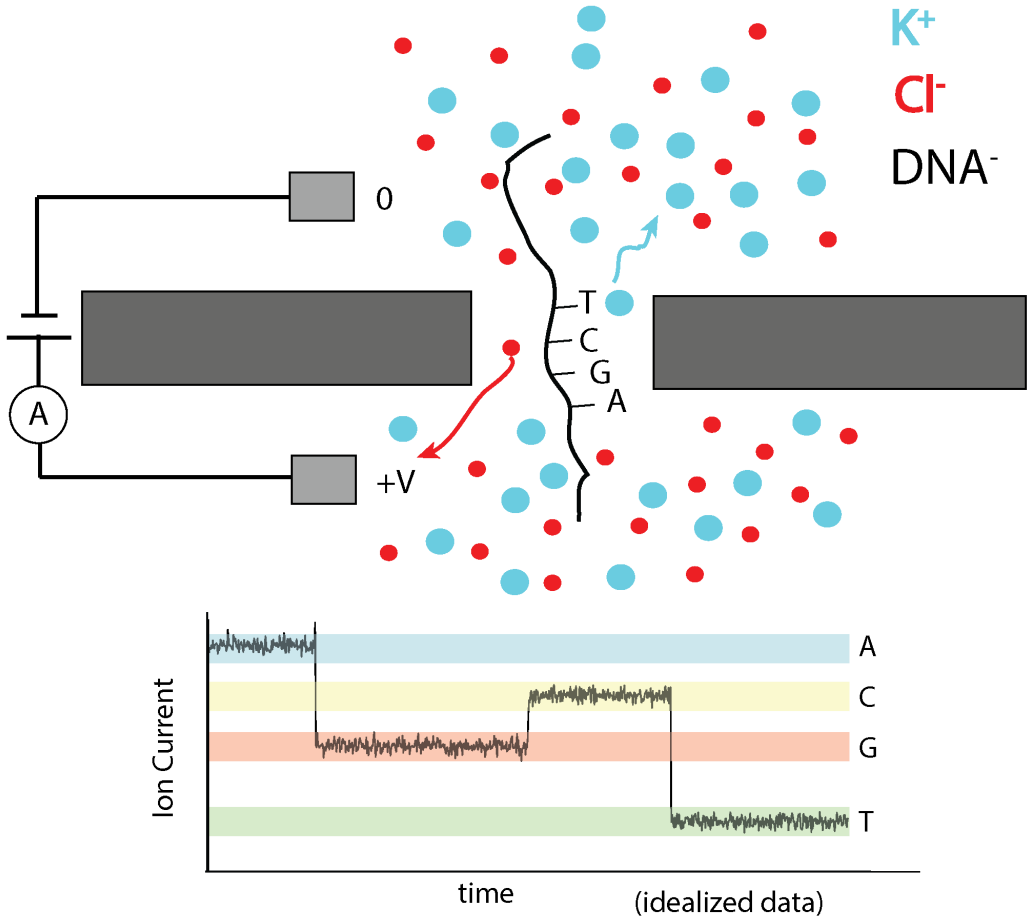

Figure 1.2: Concept of Nanopore DNA Sequencing

(Top) Nanopore DNA sequencing concept. A voltage applied across a nanopore in a membrane causes an ion current to flow through the pore. Negatively charged ssDNA is drawn into the pore by the voltage, blocking the ion current through the pore depending on the DNA bases in the pore. (Bottom) Idealized ion current trace. DNA bases are identified by a sequence of ion-current amplitudes.

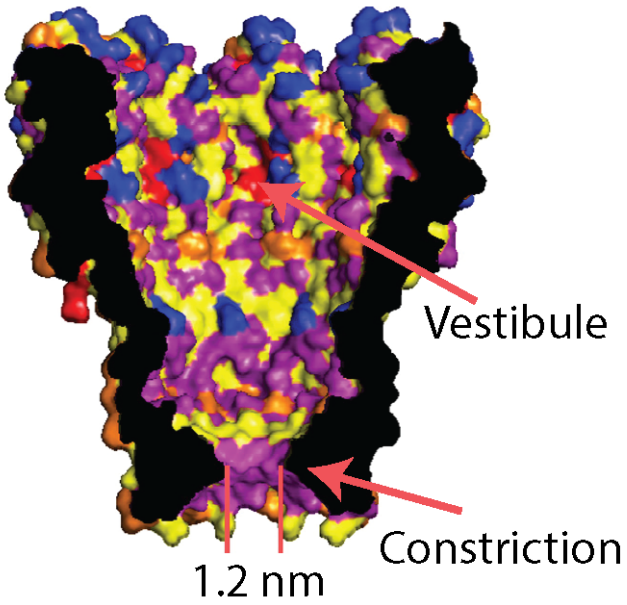

Figure 1.3: MspA nanopore

Crystal structure of the MspA nanopore [28] in the space-filling representation. This figure is modified from [27]. Colors correspond to amino-acid classes: negatively charged (blue), positively charged (red), polar (purple), non-polar aromatic (orange), non-polar aliphatic(yellow).

reproducible (figure 1.4), demonstrating the potential of the technique. Importantly, there are more than four discrete ion-current amplitudes, with about four nucleotides contributing to any given ion-current amplitude. This means that rather than having a library of four ion-current states corresponding to the four DNA bases, there is a library of 256 ion-current states corresponding to each four-letter combination of A,C,G and T [32].

Since the development of enzyme-controlled nanopore DNA sequencing, nanopore DNA sequencing with MspA has been shown to be capable of directly detecting epigenetic modifications to DNA[33, 34, 35], small chemical modifications such as the addition of a methyl group to cytosine which are believed to play a key role in gene expression [36], obtaining long DNA reads of genomic DNA [32, 37], and high quality genome scaffolding and species identification [32].

The nanopore sequencing question can be effectively phrased as: given a series of enzyme-controlled ion-current amplitudes, what is the underlying DNA sequence? In this thesis I focus on the reverse question: given a known DNA sequence, and a series of stochastic ion-current amplitudes, what can we learn about the motor enzyme that is controlling DNA translocation?

1.3 Enzyme Kinetics

Analyzing how enzymes function has been done classically through the field of chemical kinetics, which I review briefly here. Because chemical processes are random, individual events are distributed randomly in time. To describe these processes we use diagrams like the one shown in figure 1.5, which is read as molecule A decays to molecule B with mean rate k. This allows us write the following differential equation:

$$\frac{d[A]}{dt} = -\frac{d[B]}{dt} = -k \cdot [A] \tag{1.1}$$

where the notation $[A]$ stands for 'concentration of species A', which has units of molarity, or mol per liter. Equation 1.1 is called a 'rate equation'. The goal of chemical kinetics is to determine the underlying 'mechanism' or 'reaction network' of a chemical process, given

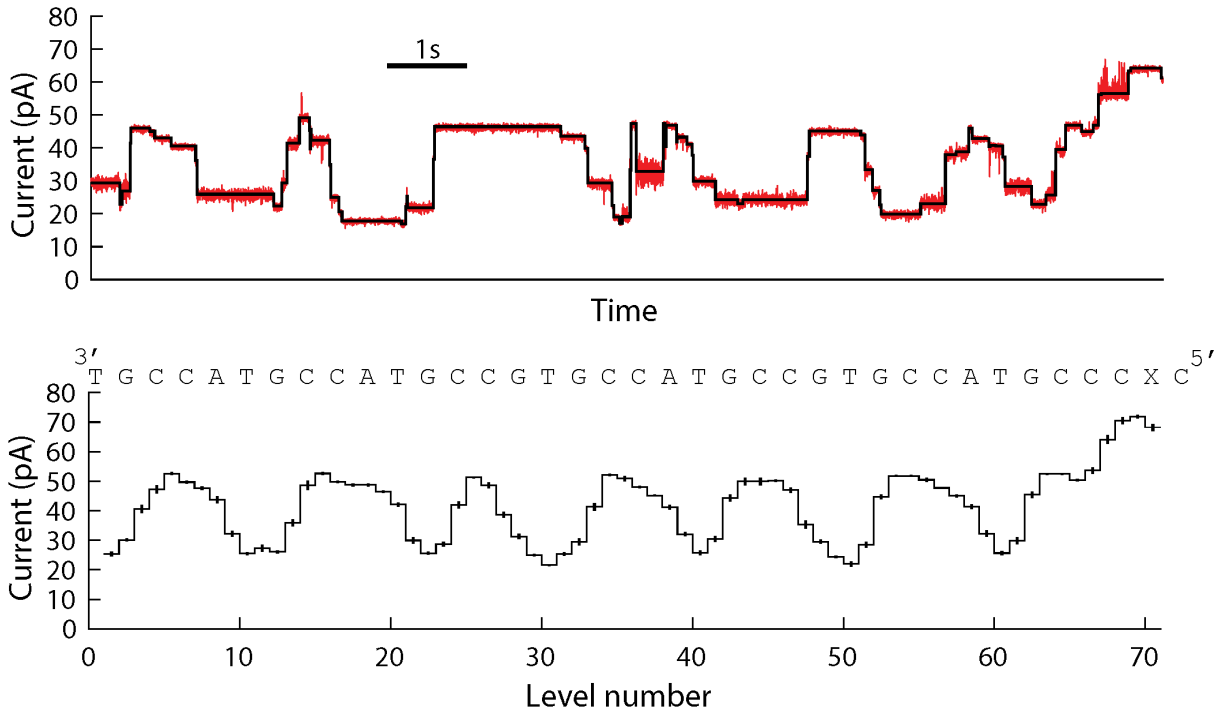

Figure 1.4: Reproducibility of Nanopore Sequencing Reads

(top) A raw ion-current versus time trace of DNA controlled by a Hel308 helicase. (bottom) A consensus of ion current reads like those from above, with the temporal information removed. The associated DNA sequence is displayed above. The 'X' in the DNA sequence indicates an abasic residue, a sugar-phosphate backbone with no attached base. Abasics are frequently used in nanopore experiements for calibration. Error bars are the standard error, demonstrating high confidence in ion-current states.

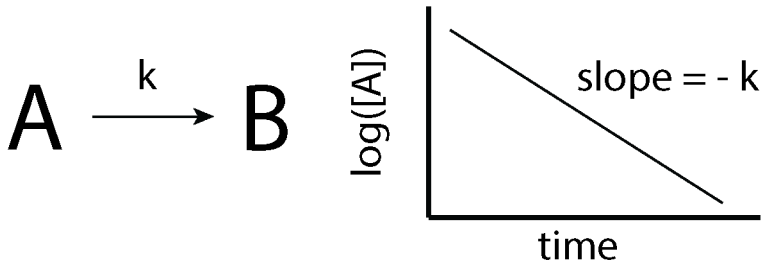

Figure 1.5: Basic Chemical Reaction Diagram

(Left) A molecule 'A' decays to a molecule 'B' with a mean rate k. (right) The concentration of species A versus time for the reaction scheme on the left. The y-axis is logarithmic.

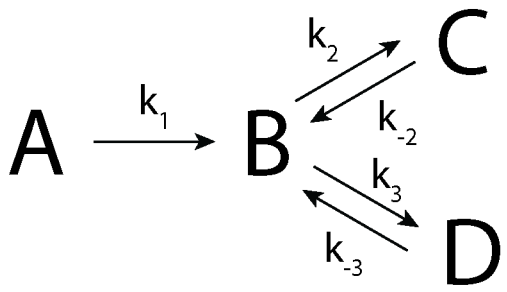

Figure 1.6: Reaction Network Diagram

A more complex reaction network. Molecule A decays to molecule B which reversibly converts between molecules C and D.

a set of experimental observables (for example, the concentration of a given molecule as a function of time). Reaction networks can be analyzed mathematically by solving a system of rate equations. The change in concentration per unit time of a given chemical species is equal to the difference between the rates 'flowing' into the sum of the and out of each state. For example, in the hypothetical model in figure 1.6 the change in concentration per unit time of state B is:

$$\frac{d[B]}{dt} = (k_1[A] + k_{-2}[C] + k_{-3}[D]) - (k_2 + k_3)[B] \tag{1.2}$$

Applying the methods of chemical kinetics discussed above, we can develop a basic mathematical model to study enzymes. Most enzyme reactions can be described as the binding of a substrate molecule, S, to an enzyme, E, which is then converted to a product molecule, P, followed by product release (Figure 1.7). Assuming that the product release is instantaneous ($k_3 \gg$ all other rates), we can write down the following system of differential equations:

$$\frac{d[E]}{dt} = -k_1 \cdot [E] \cdot [S] + k_{-1}[E \cdot S] \tag{1.3a}$$

$$\frac{d[E \cdot S]}{dt} = k_1 \cdot [E] \cdot [S] - (k_2 + k_{-1})[E \cdot S] \tag{1.3b}$$

$$\frac{d[P]}{dt} = k_2 \cdot [E \cdot S], \tag{1.3c}$$

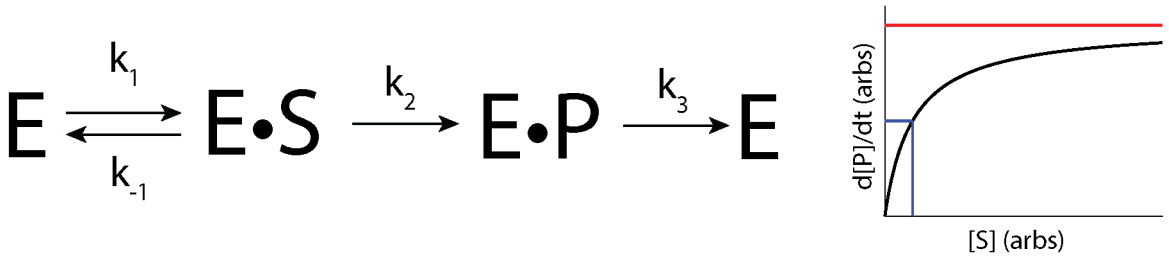

Figure 1.7: Michaelis-Menten Mechanism

(Left) The Michaelis-Menten mechanism. A substrate molecule, S, binds to an enzyme, E, which is then converted to a product molecule, P. (Right) The rate of product formation as a function of the substrate concentration. The red horizontal line indicates the value of V_m, while the vertical blue line indicates the value of K_m, at which $\frac{d[P]}{dt} = V_m/2$.

where $[E{\cdot}S]$ is the concentration of bound enzyme and $[E]$ is the concentration of free enzyme. Equation 1.3c describes the rate of production of the product molecule, which is typically the quantity that is measured in bulk experiments. Because the enzyme and substrate must come together in solution, there is a term proportional to $[E] \cdot [S]$, where both $[E]$ and $[S]$ implicitly dependent on time. This system of differential equations is therefore non-linear, and does not admit analytic solutions. Equations 1.3a-c can be solved using several different approximations [38, 39] to give:

$$\frac{d[P]}{dt} = V_m \cdot \frac{[S]}{K_m + [S]}, \tag{1.4}$$

where $V_m = k_2 \cdot [E_0]$ is the maximum rate of reaction, K_m is the substrate concentration at which the reaction rate is half of its maximum, and $[E_0]$ is the total concentration of enzyme in solution. Equation 1.4 is the well-known Michaelis-Menten equation, which describes the rate of product formation as a function of the substrate concentration. Measurements of V_m and K_m have yielded insight into the function of many enzymes [40]. However, many more complex models effectively reduce to the Michaelis-Menten equation [41], which can disguise more complex behaviors. In addition, measuring the average rate of product formation

averages over stochastic information that is inherent to enzyme kinetics [42]. In order to understand more complex enzyme mechanisms, techniques beyond bulk measurements of product formation that give access to more information about the enzyme are required.

1.4 Single-Molecule Enzyme Kinetics: Imaging One Enzyme at a Time

The solution to avoiding averaging over the stochastic behavior of enzymes is to probe single enzyme molecules. By looking at the dwell-times between successive motor-enzyme states, we can analyze the underlying probability distribution functions, as opposed to just the mean rate of product formation, yielding much more information about the underlying kinetic mechanism of the enzyme [42, 43]. In addition, by applying an external force to an enzyme, it is possible to determine how chemical energy is converted into mechanical work [41]. Many techniques have been invented to monitor the behavior of single molecules in real time, and I introduce several of the most commonly used of these briefly.

Figure 1.8 shows several illustrations of single-molecule techniques. In dual-trap optical tweezers (OT, [44]), dielectric beads are held in optical traps, with either an enzyme or DNA attached to the beads. As the length of DNA between the beads changes due to enzyme activity, the position of the beads is measured in real-time, providing a measurement of enzyme progress along the DNA. Magnetic tweezers (MT, [45]) function similarly to OT, but instead a paramagnetic bead is held in a magnetic field gradient. DNA is attached to the bead at one end and to an immovable surface at the other. As in OT, the bead position as a function of time is measured. In OT and MT a force is applied to the DNA, resulting in a force on the enzyme which can aid or hinder activity. In Förster Resonance Energy Transfer (FRET, [46]) two fluorescent labels that absorb/emit light at different frequencies are attached to biomolecules. When laser light is shined onto the system, the system will fluoresce in one color if the fluorescent labels are far apart ($\gtrsim$ 10 nm). When the labels come close together, energy is transferred from one label to the other, causing the system to fluoresce in a different color. FRET can be used to visualize the relative motion of protein domains, or the distance between a protein and a DNA substrate. Each of these approaches

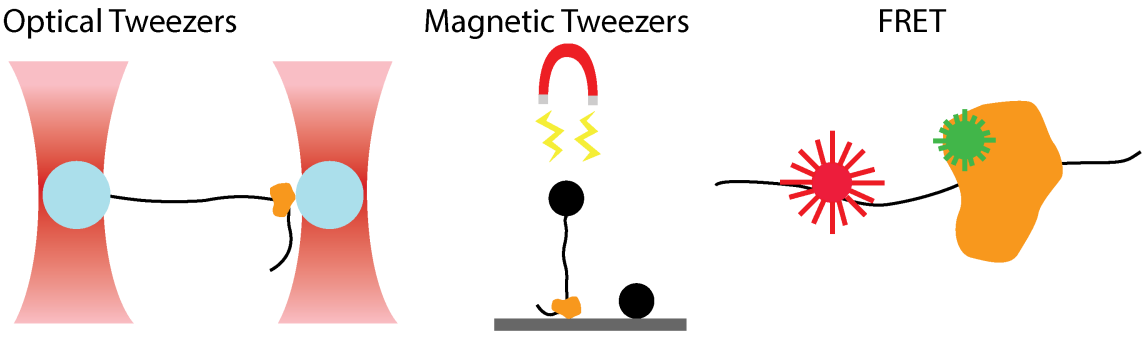

Figure 1.8: Techniques for Observing Single Enzymes

Adapted from [50]. (left) An illustration of dual-trap optical tweezers. Two dielectric beads (light blue) are held in optical traps (red). A motor enzyme (yellow) is tethered to one bead while DNA is tethered to the other (black). As the enzyme walks on the DNA, the length of DNA between the beads changes, enabling measurements of DNA position. (middle) An illustration of magnetic tweezers. An enzyme (yellow) is attached to a surface, while a bead (black ball) is held in a magnetic trap (red). As the enzyme walks the length of DNA between the surface and bead changes, enabling measurements of DNA position. A second bead is used to calibrate the relative distance between the surface and the optical trap. (right) An illustration of FRET. Two fluorescent labels (red and green) are attached to a motor enzyme (yellow). As the enzyme walks along DNA (black), the labels come closer together, leading to a change in the fluorescence signal.

has been used to provide valuable insight into the function of enzymes [47, 48, 49].

We recognized in our development of nanopore DNA sequencing that we also had a single-molecule technique for observing enzyme progress on DNA. While we had typically ignored temporal information for nanopore DNA sequencing, we were able to observe single-nucleotide phi29 DNAP steps at millisecond time scales, and in fact had a very precise record of single-enzyme movement on DNA. The goal of this thesis is to present the transformation from MspA-based nanopore DNA sequencing to using MspA for high resolution single-molecule enzyme kinetics.

In chapter 2, I discuss my work in the development of a new single-molecule technique: Single-molecule Picometer Resolution Nanopore Tweezers (SPRNT) from nanopore DNA sequencing. In chapter 3, I use the methods of single-molecule chemical kinetics and SPRNT

to analyze the behavior of the DNA helicase Hel308. In chapter 4, I extend my analysis of Hel308 to how DNA sequence regulates Hel308 kinetics. In chapter 5, I conclude the thesis with a discussion on the future of SPRNT.

Chapter 2

SUB-ANGSTROM SINGLE-MOLECULE MEASUREMENTS OF MOTOR PROTEINS USING THE MSPA NANOPORE

This work was my first major contribution to the lab, done under the guidance of Ian Derrington. An article was published on September 28, 2015 in Nature Biotechnology [51], on which I was second author. This article was our lab's first publication on SPRNT. The work in the final section of this chapter discussing the resolution of SPRNT in comparison to other single-molecule techniques was not part of this publication, and was done by Laszlo et al. in a review article for Methods on which I was not an author [50]. The discussion is included in this chapter because of its importance to SPRNT.

2.1 Introduction

The ability to directly observe the molecular motion of single molecules in real-time provides insights into enzyme function that can not be obtained by bulk assays. Before nanopores, the highest-precision single-molecule measurements had been obtained using optical tweezers, which can measure motor protein procession with $\approx$ 300 picometer (pm) spatial resolution at $\approx$ 1 second time scales [52, 46, 50]. In this chapter I present Single-molecule Picometer-Resolution Nanopore Tweezers (SPRNT), a method for monitoring the motion and conformational changes of processive nucleic-acid-binding proteins as the nucleic acid passes through a nanopore. SPRNT detects nucleic acid motion relative to the enzyme that processes it with a precision of $\approx$ 40 pm on millisecond timescales. Here we use SPRNT to observe two distinct sub-states in the ATP hydrolysis cycle of a helicase.

SPRNT is effectively the inversion of nanopore DNA sequencing [27, 31, 32, 37]. In nanopore sequencing a motor enzyme slows down DNA, enabling determination of the DNA

sequence. In this chapter, we use a known DNA sequence to make highly sensitive measurements of DNA position, and use the dwell-times of ion current levels to make kinetic measurements of the controlling motor enzyme. The basics of SPRNT are illustrated in figure 2.1. DNA bound to a motor enzyme is drawn into a single MspA by the electric field (2.1a,b). The motor enzyme then controls translocation of the DNA through the nanopore, leading to a time series of discrete current levels, which are reproducible to the picoampere scale (2.1c,d). The time domain information from 2.1c now provides kinetic measurements of the controlling motor enzyme, which can be used to infer mechanisms of enzyme motion [53].

2.2 Results

For some DNA sequence contexts there is a large change in ion current when the DNA polymerase phi29 moves the DNA by a single nucleotide. The ion current levels associated with a sequence of DNA containing an abasic site (marked by an 'X') has a change in current equal to 16 pA when the DNA moved by one nucleotide (Fig. 2.2a). If the DNA were to move within MspA by a distance of about one tenth of a nucleotide, a linear interpolation would have the observed current change by about one tenth of the change in current, or approximately 1.6 pA. Coupling the ion-current to the DNA position allows us to measure the position of DNA in MspA to precision much smaller than one nucleotide. The scale in Fig. 2.2a, which shows the conversion of current to displacement, uses a cubic spline to approximate the ion current between levels measured at 1 nt intervals. Using this distance scale we relate the uncertainty of ion current levels to the uncertainty of the DNA position in the pore using standard error propagation. For the ion current levels depicted in Fig. 2.2a a position uncertainty as small as 0.06 nt can be resolved, corresponding to a distance uncertainty of 40 pm, assuming an inter-phosphate distance to be 690 p.m [54, 55]. and 88-95% DNA-elongation.

Next, we changed the elongation of DNA by altering the electrostatic force applied to the DNA. Whilst DNA was moved by phi29 DNAP in single nucleotide steps, we applied driving

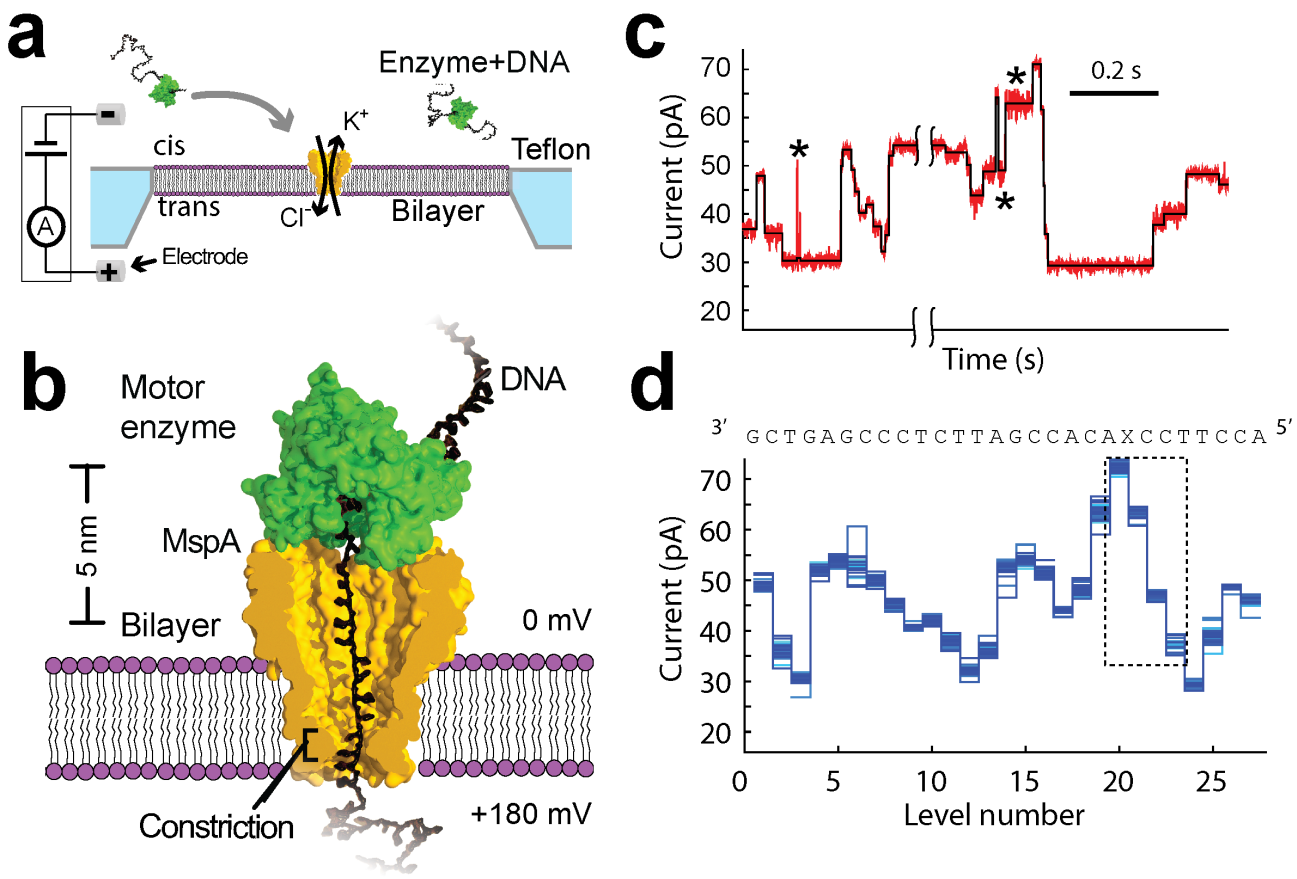

Figure 2.1: Schematic of SPRNT

(a) Schematic of the SPRNT system. Enzyme DNA complexs are drawn into the pore by the electric field. (b) ssDNA bound to the motor enzyme (polymerase or helicase) threads through the pores constriction until the enzyme comes to rest on the pore rim. The enzyme controls the DNAs motion through the pore, while the nucleotides positioned within MspAs constriction govern the ion current. (c) The phi29 DNA polymerase (DNAP) moves the DNA through MspA in single-nucleotide steps resulting in distinct current levels. Black lines mark the average current of observed levels. Breaks in the current trace are for current levels lasting more than 200 ms. Back-stepping of the phi29 DNAP causes repetitions of levels, indicated by *. (d) The mean ion current of the time-ordered levels and overlay the pattern of current levels for 31 recordings of the same sequence of DNA. The associated DNA sequence is shown; X is an abasic residue.

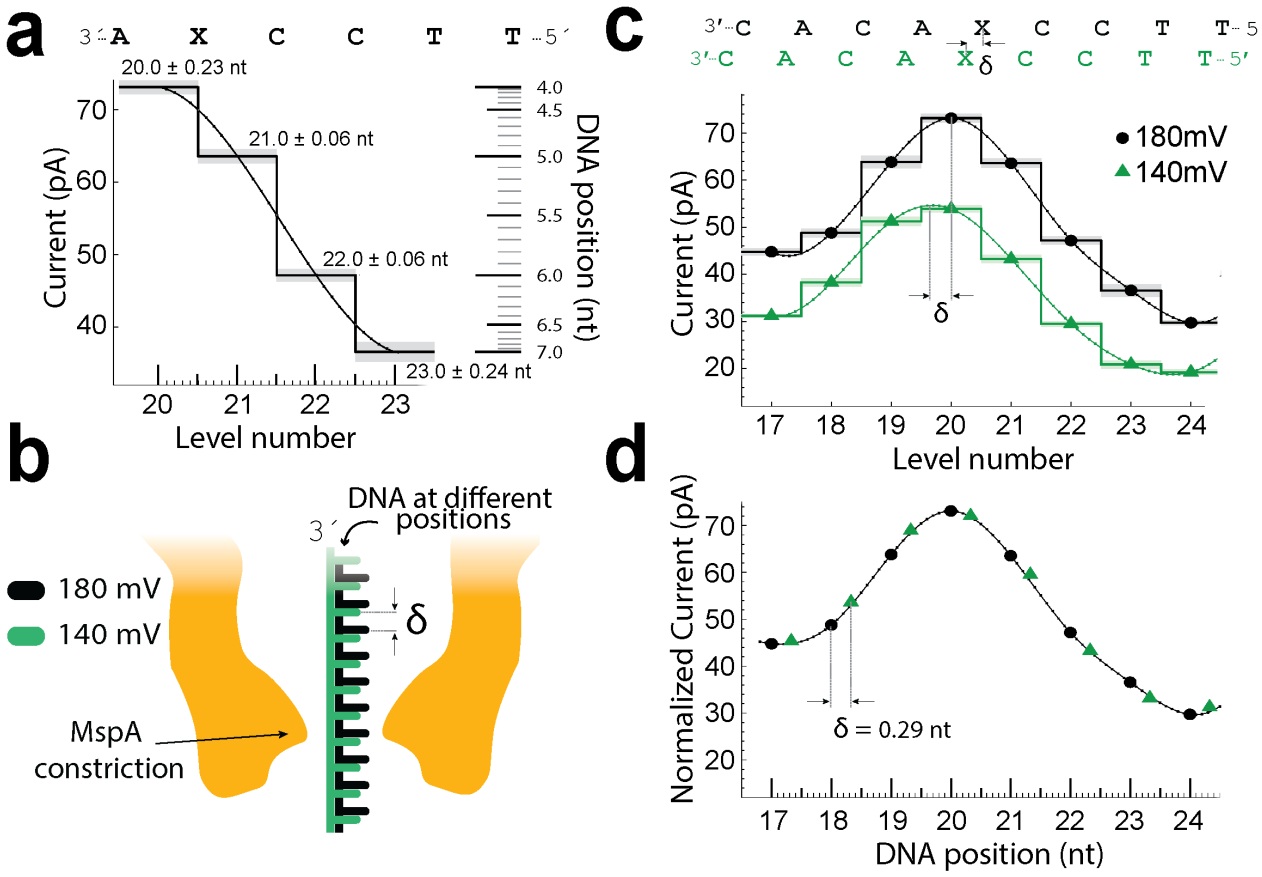

Figure 2.2: DNA Position Measurements with SPRNT

(a) Zoomed in view of the dashed box in 2.1d demonstrates conversion of current measurement to DNA position using a smooth curve (spline) fit to the current pattern. The standard deviation translates to uncertainty in DNA position is as low as 0.06 nt. (b) Illustration showing that lower voltage, i.e. decreased force, reduces the DNAs elongation and shifts its position within MspAs constriction. (c) Comparison of ion current levels recorded at 180 mV (black circles) and at 140 mV (green triangles). Peaks of the spline interpolation illustrate a shift of the DNAs position. (d) Current values for 180 mV (black circles) and a spline interpolation to those levels (black curve). Green triangles show the current levels taken at 140 mV in panel (c) after applying a multiplicative scale and additive current offset. The scaled 140 mV levels were horizontally displaced by $\delta = 0.29$ nt to put them in line with the 180 mV spline. 'Level number' refers to the number assigned to each level as it appears in order while 'DNA position' refers to the position of DNA within the pore. We define integer DNA positions to be identical to phi29 DNAP level numbers.

potentials of 140 mV and 180 mV. Changing the voltage (and thereby the force on the DNA) alters the elongation of DNA between the motor enzyme and pore constriction and shifts the position of nucleotides within MspAs constriction (Fig. 2.2b). Figure 2.2c displays the ion-current levels for data taken at the two voltages with cubic spline interpolants overlaid. The location of the splines peaks shift between the different voltages. After normalizing the current amplitudes, we find that the spline for levels taken at 180 mV can predict the levels at 140 mV, when the spline is shifted 0.29 ± 0.03 nt (Fig. 2.2c). Exploring DNA elongation with voltages between 100 mV and 200 mV indicated that the DNA elongation was consistent with experimental force-stretching curves for ssDNA [54, 55] for forces in the range of ≈20-50 pN (Appendix A.3). These results show that the spline is a reasonable prediction of currents between levels seen at 1 nt intervals.

We evaluated the precision of SPRNT using the helicase Hel308, which is an ATP-dependent Ski2-like superfamily II (SF2) helicase/translocase that unwinds duplex DNA in the $3'$ to $5'$ direction. Hel308 is conserved in many archaea and eukaryotes, including humans [56, 57, 58, 59]. With a known crystal structure, Hel308 is a good system for understanding processive SF2 helicases [60]. We used Hel308 of *Thermococcus gammatolerans* EJ3 (Accession number: YP_002959236.1) (hereafter Hel308). The current patterns we observed were qualitatively similar to those observed with phi29 DNAP (Fig 2.3a,b). However, when Hel308 moved DNA through the pore, we observed nearly twice the number of levels as compared to when phi29 DNAP moved DNA through the pore, even though the same length of DNA passed through the pore (2.3a,b).

By comparing 72 Hel308 DNA translocation events, we produced a consensus set of current levels for Hel308 DNA translocations of DNA 'sequence A' through the pore (Fig. 2.3c,d, Appendix A.4, Fig. 2.4,2.5). We used this consensus set to deduce the position of DNA when Hel308 controlled DNA translocation compared with the position of the same DNA sequence moved by phi29 DNAP (Fig. 2.3a). We found that the odd numbered Hel308 current levels correspond to the DNA being held 0.14 ± 0.03 nt higher in the pore than the closest corresponding current level taken with phi29 DNAP. We found the even numbered

Hel308 levels correspond to the DNA being held 0.41 ± 0.03 nt lower in the pore. The average difference in position between the odd and even numbered Hel308 steps is therefore 0.55 ± 0.04 nt (Fig. A.5).

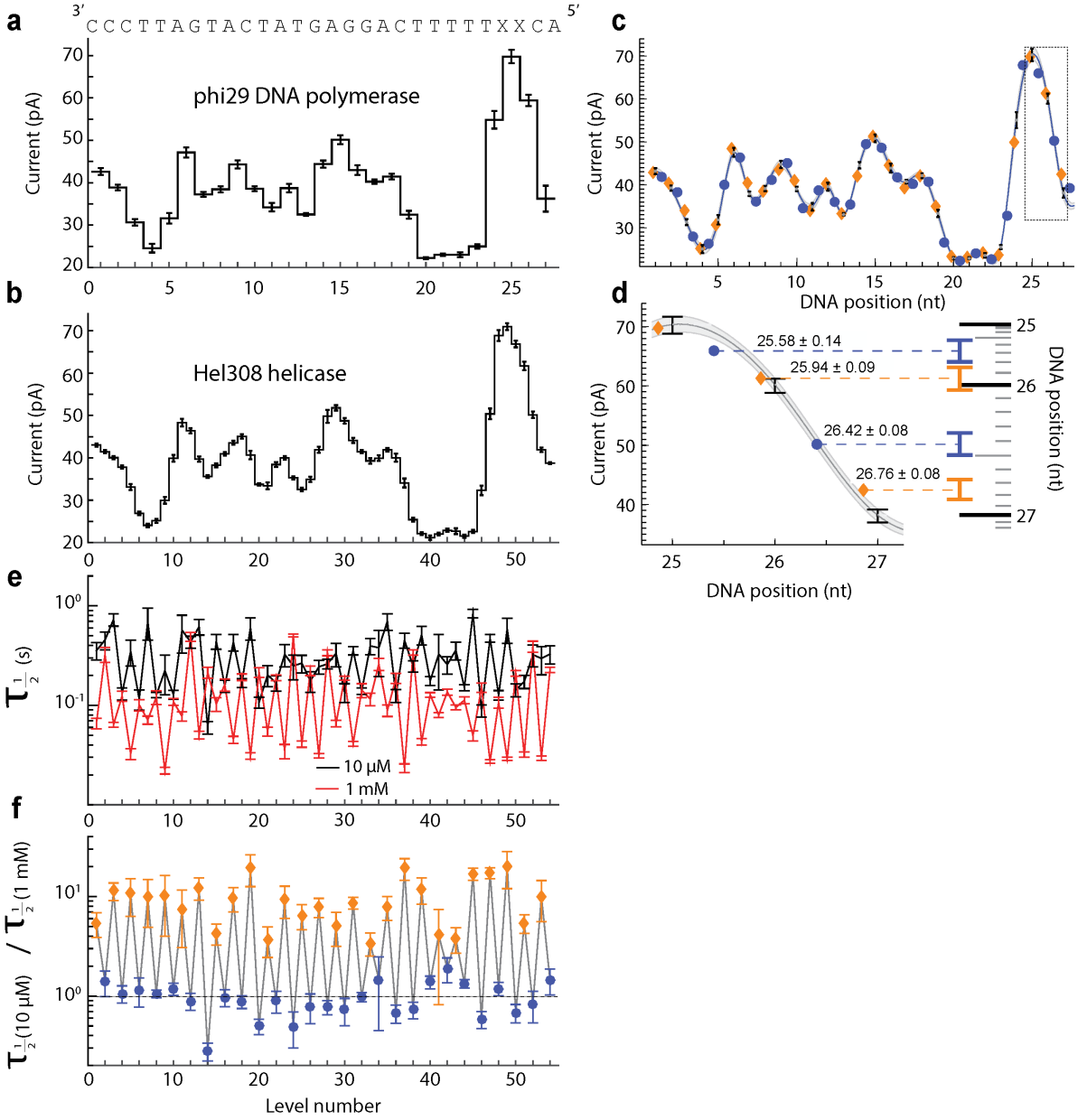
a
3′ C C C T T A G T A C T A T G A G G A C T T T T T X X C A 5′
phi29 DNA polymerase
Current (pA)
b
Hel308 helicase
Current (pA)
c
Current (pA)
DNA position (nt)
d
25.58 ± 0.14
25.94 ± 0.09
26.42 ± 0.08
26.76 ± 0.08
Current (pA)
DNA position (nt)
DNA position (nt)
e
τ1/2 (s)
10 μM
1 mM
f
τ1/2(10 μM) / τ1/2(1 mM)
Level number

Figure 2.3: SPRNT Analysis of Hel308 DNA Helicase

(a) Consensus of current level patters for 20 reads of DNA 'sequence A' with phi29 DNAP controlling DNA translocation through the pore. (b) Same as in a except for 72 reads using Hel308 to control the DNA motion. (c) Means of ion current levels recorded with Hel308 actuated DNA movement (orange and blue symbols) scaled to match a spline (grey curve) of the levels found with phi29 DNAP-controlled movement (black points), also shown in (a). The shaded levels in (b), indicated with orange diamonds, were similar to levels found with phi29 DNAP but were horizontally offset by -0.14 nt in order to best match the spline of levels taken with phi29 DNAP. The unshaded levels in (b), indicated with blue circles, were offset by + 0.41 nt relative to the single nucleotide step positions taken by phi29 DNAP. (d) Expanded view of DNA positions 25 through 27, with colors indicating the same elements as (c). As in Figures 2.2a,c, we illustrate the use of the spline of the phi29-DNAP levels as a distance scale to find the position of even and odd numbered levels found with Hel308 (Appendix A.5). (e) Median duration of corresponding current levels in (b) for two different ATP concentrations: 10 μM (blue) and 1 mM (red). (f) The ratio of the median durations with high and low [ATP] removes sequence dependence that also influences the step durations. The levels alternate between ATP-independent levels (marked with blue dots) and ATP-dependent levels (marked with orange diamonds).

Next, we examined the median duration ($\tau_{1/2}$) of each level at different ATP concentrations (Fig. 2.3e). In Fig. 2.3f we compare the median duration of each current level at 10 μM ATP to those at 1 mM ATP by dividing $\tau_{1/2}(10\ \mu M)$ by $\tau_{1/2}(1\ \text{mM})$. We found that the durations for even numbered levels depend on [ATP] while durations for odd numbered levels are independent of [ATP]. The ion current magnitude did not change with [ATP]. A full [ATP] titration is described and shown in Appendix A.6 and figure A.6, and is discussed in detail in chapter 3.

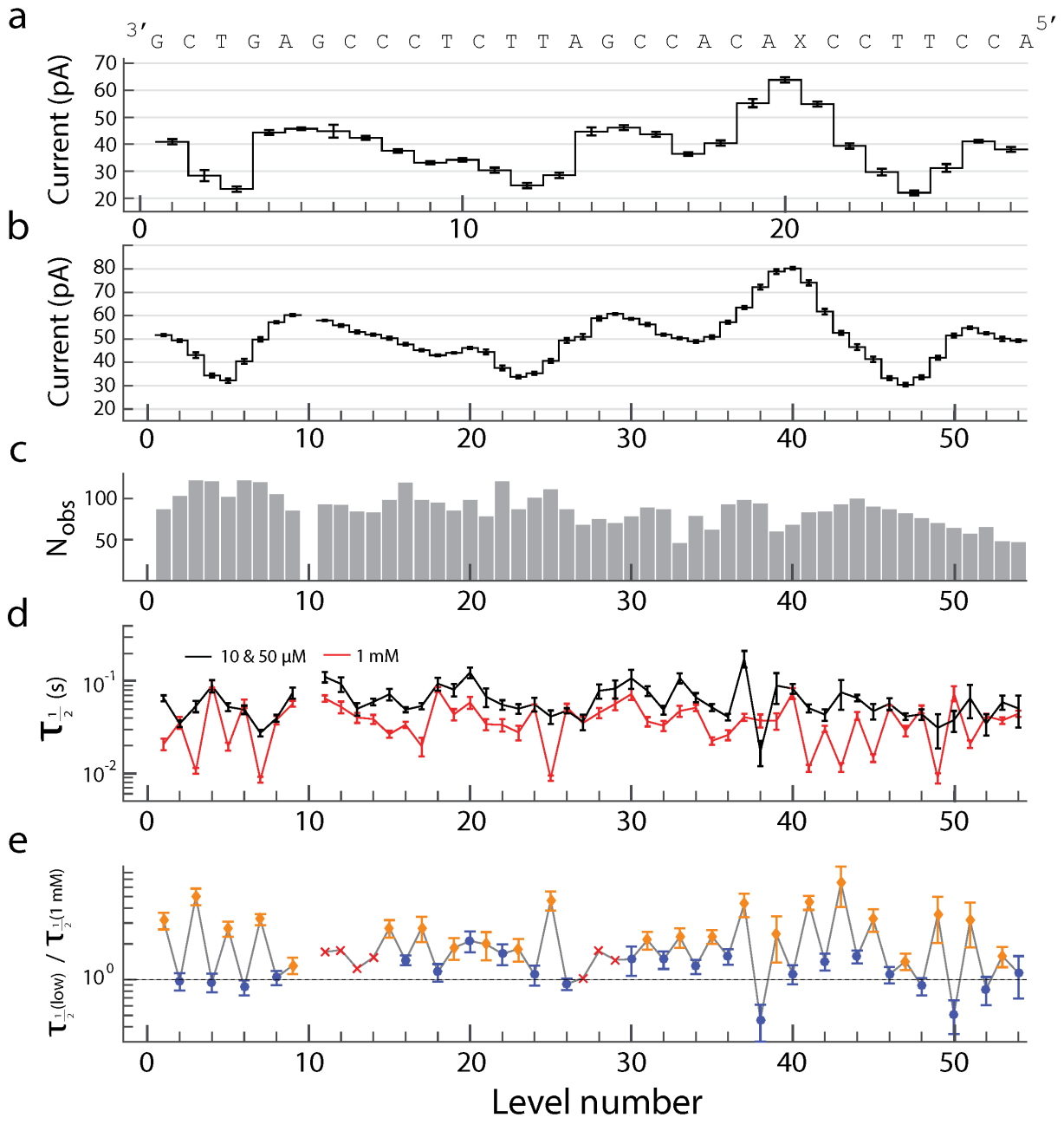
a
3′ G C T G A G C C C T C T T A G C C A C A X C C T T C C A 5′
Current (pA)
b
Current (pA)
c
N_obs
d
10 & 50 µM
1 mM
τ½ (s)
e
τ½(low) / τ½(1 mM)
Level number

Figure 2.4: Sequence B Consensus

A combination of information in figure 2.3 but for DNA sequence B. (a) The observed level patterns for phi29 DNAP moving sequence B through MspA. Data was taken with 150 mM [KCl] in the cis well and 500 mM [KCl] in the trans well. (b) The observed level patterns for Hel308 translocating DNA sequence B through MspA (black lines). The automatically generated consensus levels for sequence B are aligned to the sequence and level pattern found for phi29 DNAP. Data was taken with 400 mM [KCl] buffers in both cis and trans wells. The difference in salt conditions accounts for the difference in current values between (a) and (b). A gap indicates a position where a level was missing due to degeneracy. (c) The number of times that a given level shown in (b) was observed. (d) The median duration of the levels with the current shown in (b) while using 10 and 50 μM [ATP] (black line) and using 1 mM [ATP] (red line). Level durations depend partially on sequence context. (e) Ratio of level durations for observations using low [ATP] (10 and 50 μM) and for high [ATP] (1 mM). We indicate odd-numbered levels that depend on [ATP] with an orange diamond and even-numbered levels that do not depend on [ATP] with a blue circle. Levels that could not be identified as ATP-dependent or independent due to degeneracy of nearby current values are indicated with a red ‘x‘. Comparing the duration at different values of [ATP] removes level-duration dependence on sequence context.

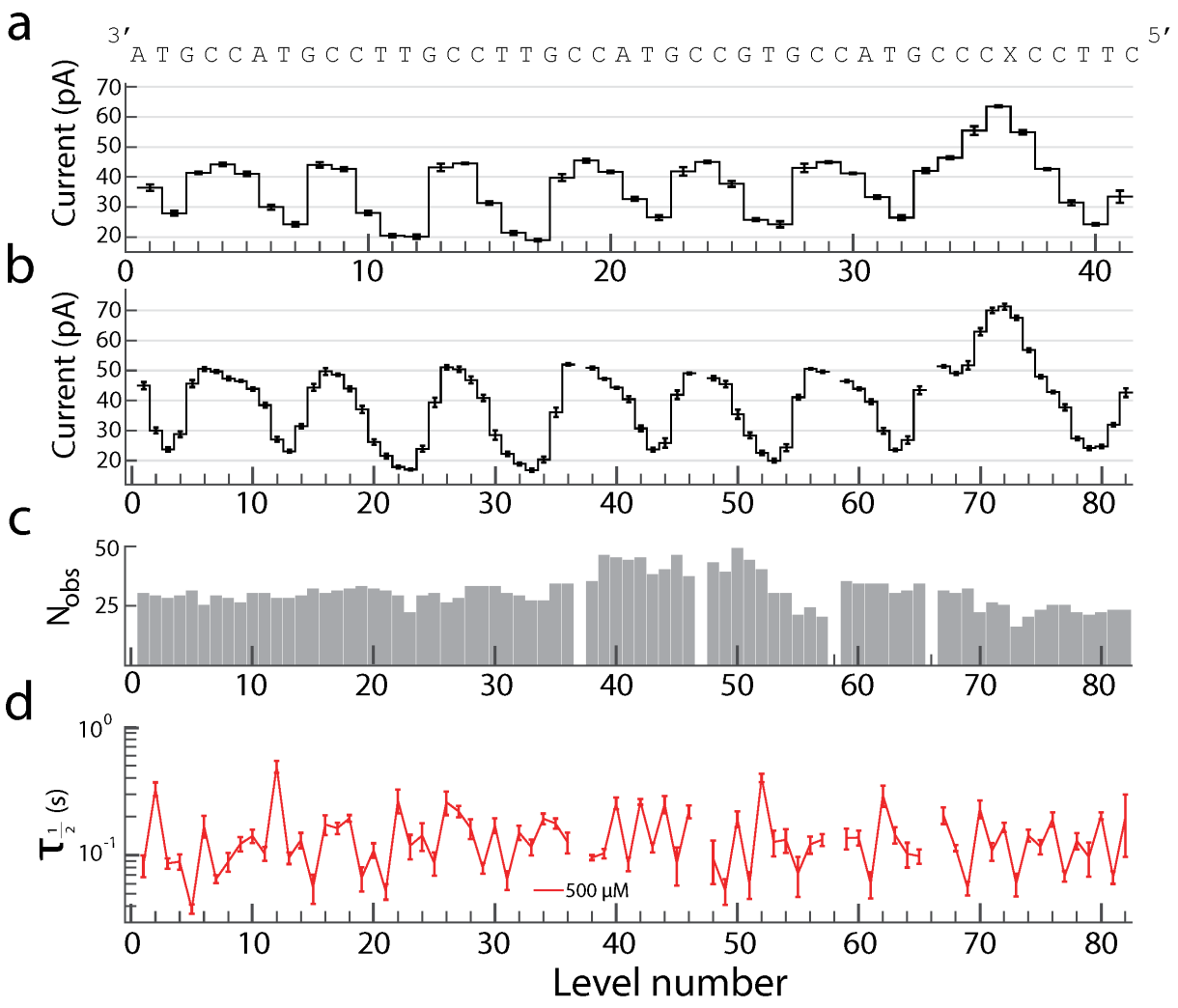
a
3′ A T G C C A T G C C T T G C C T T G C C A T G C C G T G C C A T G C C C X C C T T C 5′
Current (pA)
b
Current (pA)
c
N_obs
d
τ½ (s)
500 µM
Level number

Figure 2.5: Sequence C Consensus

A combination of information in figure 2.3 but for DNA sequence C. (a) The observed level patterns for phi29 DNAP moving sequence B through MspA. The sequence was designed to have repeating pattern with high contrast between adjacent current levels. Data was taken with 150 mM [KCl] in the cis well and 500 mM [KCl] in the trans well. (b) The observed level patterns for Hel308 translocating DNA sequence C through MspA (black lines). The automatically generated consensus levels for sequence C are aligned to the sequence and level pattern found for phi29 DNAP. Data was taken with 400 mM [KCl] buffers in both cis and trans wells. The difference in salt conditions accounts for the difference in current values between (a) and (b). A gap indicates a position where a level was missing due to degeneracy. (c) The number of times that a given level shown in (b) was observed. (d) The median duration of the current levels shown in (b) while using 500 μM [ATP]. We identified odd-numbered levels as ATP-dependent and even-numbered levels as ATP-independent levels for sequence C based on the results of figure 2.3c, indicating that ATP independent levels will generally have a longer median duration at 500 μM [ATP].

2.3 Discussion

Büttner et al.'s analysis of the crystal structures of Hel308 and the SF2 helicase, Vasa, revealed large conformational shifts upon ATP binding [60]. Büttner et al. propose an inchworm model in which the two RecA-fold DNA binding domains, through the action of ATP binding and then hydrolysis, take turns moving along a DNA strand. In SPRNT the movement of the DNA in the MspA pore is likely a combination of the movement of DNA inside Hel308 and conformational changes of the Hel308 that reposition Hel308 on the MspA rim (thereby changing the position of DNA inside the pore; Fig. A.7). Even so, our observations seem to confirm the model predicted by Büttner et al. Using Büttners model [60], we suggest that motif IV within domain 2 pushes the DNA upwards toward domain 1 upon ATP binding, thereby pushing the DNA partially upward within the pore (Fig. A.7b,c). ATP hydrolysis and ADP release finishes the hydrolysis cycle advancing the DNA and finishing the single nucleotide step. Previously, sub-state kinetic steps have only been inferred indirectly through fitting of durations in helicase systems [61]. However, to our knowledge, no other real-time single molecule method has allowed direct observation of sub-states within individual hydrolysis cycles of helicase kinetics.

To maximize SPRNTs resolution, it is important to choose DNA sequences that produce current levels with large differences (i.e. not homopolymeric sequences). The current between full nucleotide steps may differ from the spline interpolation that we used. Finally, during SPRNT, MspA is in contact with the enzyme and applies a 20 to 50 pN force to the enzyme (Appendix A.3). These forces and contact with MspA may alter the enzymes activity.

In addition to sub-angstrom resolution, SPRNT simultaneously provides the exact location of the enzyme along the DNA sequence via nanopore DNA sequencing [32]. This means that SPRNT could be used to answer important questions in many motor enzyme systems such as how nucleic acid sequence and structure relate to pausing and other motor enzyme activity [62]. SPRNT can resolve smaller motions of enzyme subdomains than FRET and could be used with DNA and RNA polymerases or translocases, a ribosome, or transcription

complexes. Other potential applications include analyzing reactive molecules tethered to a polymer (DNA, RNA, or hybrids) that is held in the pore.

2.4 Spatiotemporal Resolution of SPRNT

SPRNT, Optical tweezers, Magnetic Tweezers and FRET each seek to resolve transitions between discrete steps in noisy data. The signal to noise ratio (SNR) is the mean difference between two measurements divided by the noise in the measurements. The SNR quantifies how well two steps can be distinguished, and is therefore the quantity of interest when comparing different measurement techniques. Assuming that the noise in each step is Gaussian distributed, the SNR is given by (figure 2.6):

$$SNR = \frac{\Delta x}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{2.1}$$

where Δx is the difference between the mean of each step, and σ_1 and σ_2 are the noise in each measurement. The larger the SNR, the more easily two steps can be distinguished. The SNR can be increased by averaging over the position-space data, but temporal resolution is lost by the averaging process. This trade-off means that spatial and temporal resolution are fundamentally coupled quantities, as can be summarized by the following relation [50]:

$$\Delta x \cdot \sqrt{\Delta t} \geq C \cdot SNR \tag{2.2}$$

where C is a constant that depends on experimental conditions, Δx is the step size and Δt is the measurement time. To compare experimental techniques to one another, the scaled SNR ($sSNR$) is defined as the SNR at 1 second observation time ($\Delta t = 1\ s$) and 1 nanometer step size ($\Delta x = 1\ nm$). The $sSNR$ for SPRNT is more than 50 times larger than OT ([50],2.1). This information is summarized in the sensitivity plot in figure 2.7. The diagonal lines shown for each technique show the tradeoff between spatial and temporal resolution. SPRNT owes its superior resolution to the fact that the measurement of DNA position is made much closer to the enzyme than in OT or MT. In SPRNT the distance between the

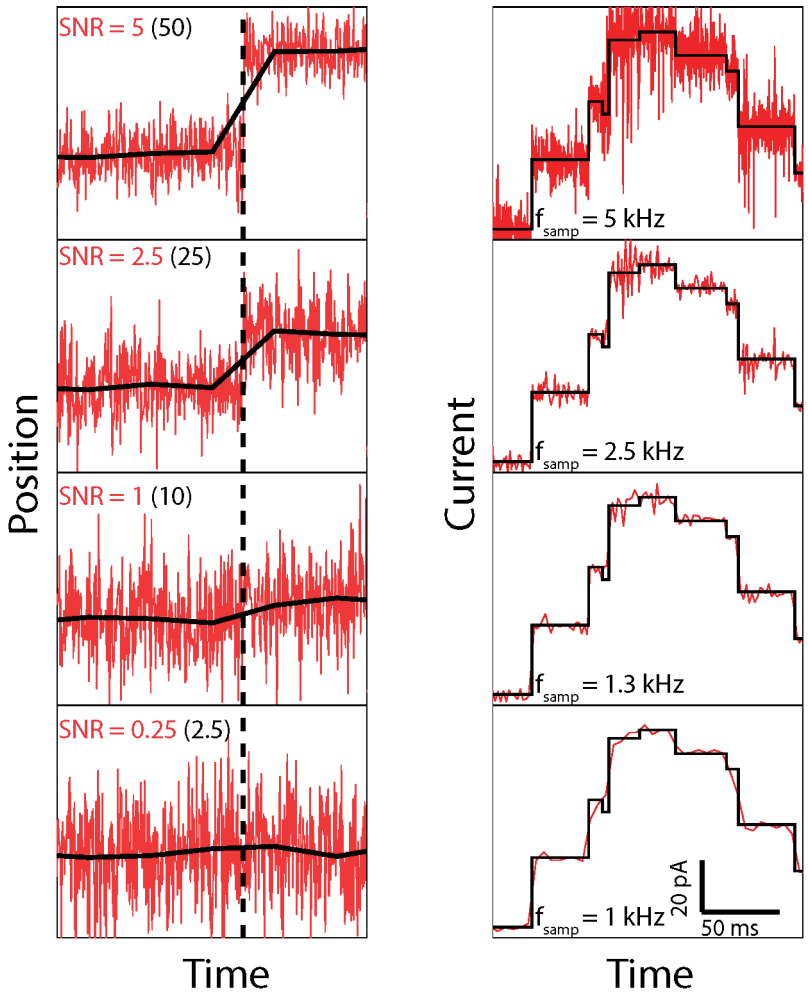

Figure 2.6: Resolving Steps in Noisy Data

(left, red) Computer-generated noisy position versus time traces. Noise is Gaussian distributed. (black) The data, reduced by averaging together every 100 data points. The signal is less noisy, but the number of samples is decreased. The black dashed line shows the location of a transition between steps. (right, red) Ion current versus time nanopore data taken with the Hel308 helicase. (black) Automatically found ion-current states. The data is downsampled by averaging further in each subsequent panel. At 5 kHz clear steps are resolved. In the bottom panel the data has been downsampled by averaging to 1 kHz (5 ms averaging time), leading to a reduction in the noise, but many Hel308 steps are too short to be seen at this sampling frequency.

Technique	sSNR	Force(pN)	Distance range (nM)	torque?	Massively Parallelizable ?
SPRNT	2360	15-60	0.04 - 10^5	No	Yes
OT	41.6	0.1 - 100	0.1 - 10^5	Yes	No
MT	24.3	0.001 - 10000	0.5 - 10^5	Yes	Yes
TIR-FRET	41.6	–	2 - 10	No	Yes

Table 2.1: Comparing Single-molecule Techniques

Adopted from [50]. A comparison of the properties of several single-molecule techniques

enzyme and measurement position is $\approx 10\ nm$ where as in OT the distance between beads is $\approx 1.5\ \mu m$. If we treat the DNA as a simple entropic spring, then the fluctuations in DNA position grow linearly with the number of links in the chain $\langle \Delta x^2 \rangle \approx N$. N itself grows linearly with the DNA length, so we can use equation 2.1 to estimate that for the same step size Δx, $\frac{SNR_{SPRNT}}{SNR_{OT}} \approx \frac{\sigma_{OT}}{\sigma_{SPRNT}} \approx \sqrt{\frac{N_{OT}}{N_{SPRNT}}} \approx \sqrt{\frac{l_{OT}}{l_{SPRNT}}} \approx \sqrt{\frac{1500}{10}} \approx 12$. This order-of-magnitude estimate provides an intuitive idea for why SPRNT has a much higher SNR than OT.

It is important to note that in SPRNT, OT, and MT that the spatiotemporal resolution is affected by the applied force. For example, because the noise in OT is Brownian motion limited, and the $sSNR_{OT}$ quoted here is at 10 pN of applied force, if we were to increase the force of the optical trap to 35 pN, $sSNR_{OT}$ would increase by a factor of 3.5. Such considerations may become important when choosing a single molecule technique for an experiment, because certain enzymes can stall at high forces [63, 64, 65, 66]. Other considerations may be important as well. For example, rotary motors such as F_1-ATP synthase can be mechanochemically probed through the application of an external torque [67], which SPRNT cannot provide. Table 2.1 summarizes properties of several single-molecule techniques.

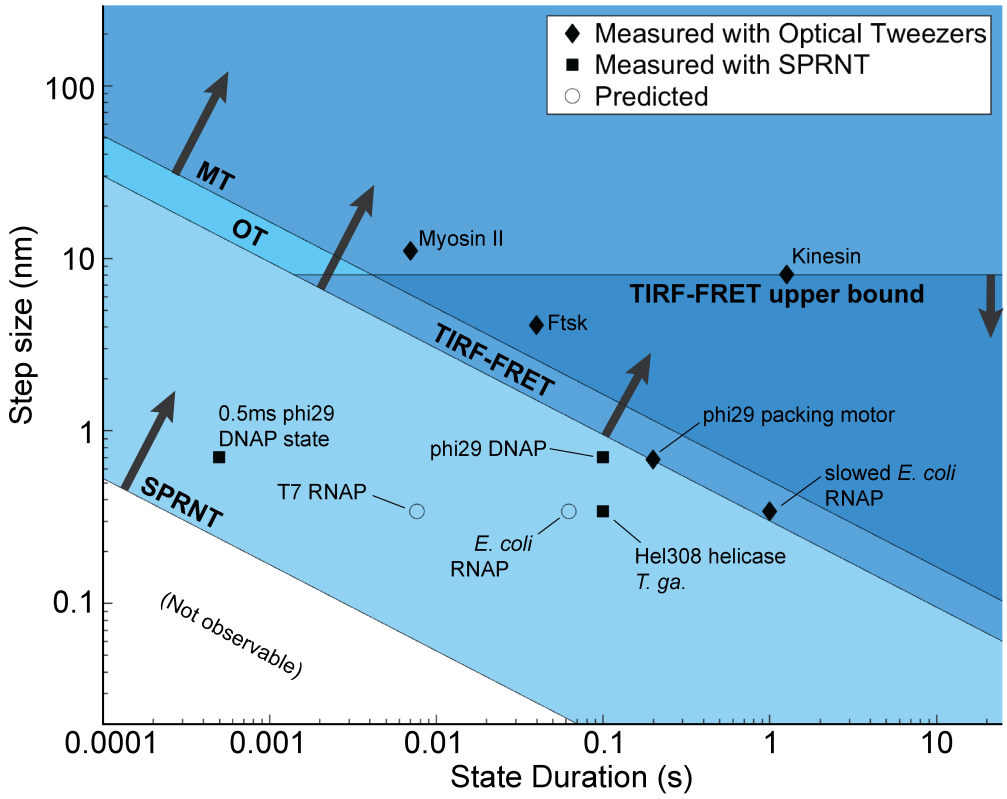

Figure 2.7: Comparing Spatiotemporal Resolution of Single-molecule Techniques

Adapted from [50], comparing the spatiotemporal resolution of various single-molecule techniques. The y-axis is the smallest step size that can be measured for a given state duration. Arrows indicate in what spatiotemporal range a given experimental technique can measure. For example, only SPRNT can measure a 1 nt step (≈0.6 nm) at 1 ms resolution.

Chapter 3

REVEALING THE KINETIC MECHANISM OF A SF2 HELICASE USING SPRNT

This work expands on the results of the previous chapter by analyzing Hel308 stepping behavior in high detail. Because SPRNT has exquisite spatiotemporal resolution, we are able to examine the kinetics of Hel308 in ways that have not been done before. The measurements done in this chapter cannot be done with any other technique.

3.1 Introduction

Enzymes, such as helicases, polymerases, translocases, and ribosomes that move along DNA or RNA perform the core functions of replication and expression in all of biology. Helicases are molecular motors that catalyze the unwinding of double-stranded DNA or RNA powered by ATP hydrolysis. Due to helicases vital role in genome maintenance, helicase defects are specifically linked to various cancers [68, 69, 70], and aging disorders[71]. Helicases are divided into six superfamilies (SF) based on structure and function. Superfamilies SF1 and SF2 are comprised of the monomeric helicases[11, 72, 73]. Structural studies of SF1 and SF2 helicases have revealed many conserved residues and motifs involved in walking along DNA and ATP binding and hydrolysis. Hel308 is a representative of the superfamily 2A DNA helicases/translocases, and possesses structure that is highly conserved in both archaea and eukarya, including humans[59, 74]. It has been proposed that Hel308 is recruited to stalled replication forks to restart the replication process[74, 56]. Hel308 is an interesting system because it requires the coordination of several protein domains in addition to the walker motifs[59] that are ubiquitous in SF1 and SF2 helicases, and can be used to develop general models for how SF1 and SF2 helicases move along ssDNA. Hel308 has been primarily studied

by bulk assays, which cannot directly probe the mechanisms by which it walks on and unwinds DNA.

Single-molecule technologies that monitor the kinetics of single enzymes at high resolution in real time have enhanced mechanistic understanding of helicases and other motor enzymes. Techniques such as optical tweezers [44], magnetic tweezers [45], and Frster resonance energy transfer[75] have been used to infer kinetic mechanisms of helicases such as UvrD [76, 77], PcrA [78] , Hepatitis C NS3 helicase [61, 79], RecQ [80], and XPD-like helicases [81]. It has been shown that SFI and SFII helicases step in single-nucleotide steps and that ATP binding causes a conformational change followed by ATP hydrolysis, ADP release, and a conformational change back to the original state resulting in a single nucleotide step along the DNA [77, 78, 79, 81, 82]. During this cycle, changes in how tightly the two recA-like domains hold onto the DNA backbone enable processive, inchworm-like motion of the helicase along the DNA. While much is known, the exact timing and choreography of these events is unclear[11]. To fully understand the exact mechanism by which ATP hydrolysis coordinates the directed motion of the helicase along DNA, a technique with the ability to resolve kinetic substeps of the hydrolysis cycle is required.

In the last chapter, we directly observed that the Hel308 helicase from *Thermococcus Gammatollerans* takes two half-nucleotide steps per nucleotide translocated along ssDNA [51] (Fig. 3.1a,b). Increasing [ATP] caused the average duration of steps at half-integer nucleotide positions to decrease, while the average duration of steps at integer nucleotide positions did not change, demonstrating the presence of two observable substates of the Hel308 ATP hydrolysis cycle (Fig. 3.1c). We call the half-integer DNA positions '[ATP]-dependent steps' and the integer DNA positions '[ATP]-independent steps'. Here, we use SPRNT to examine thousands of reads of Hel308 translocation (Table B.1) on ssDNA with two major goals: first, to develop a kinetic model that places known chemical processes such as ATP binding, ATP hydrolysis, and ADP release in the context of the mechanical motion of the enzyme observed using SPRNT. Second, we analyze the dwell times at each DNA position separately to look for sequence-dependent translocation of Hel308.

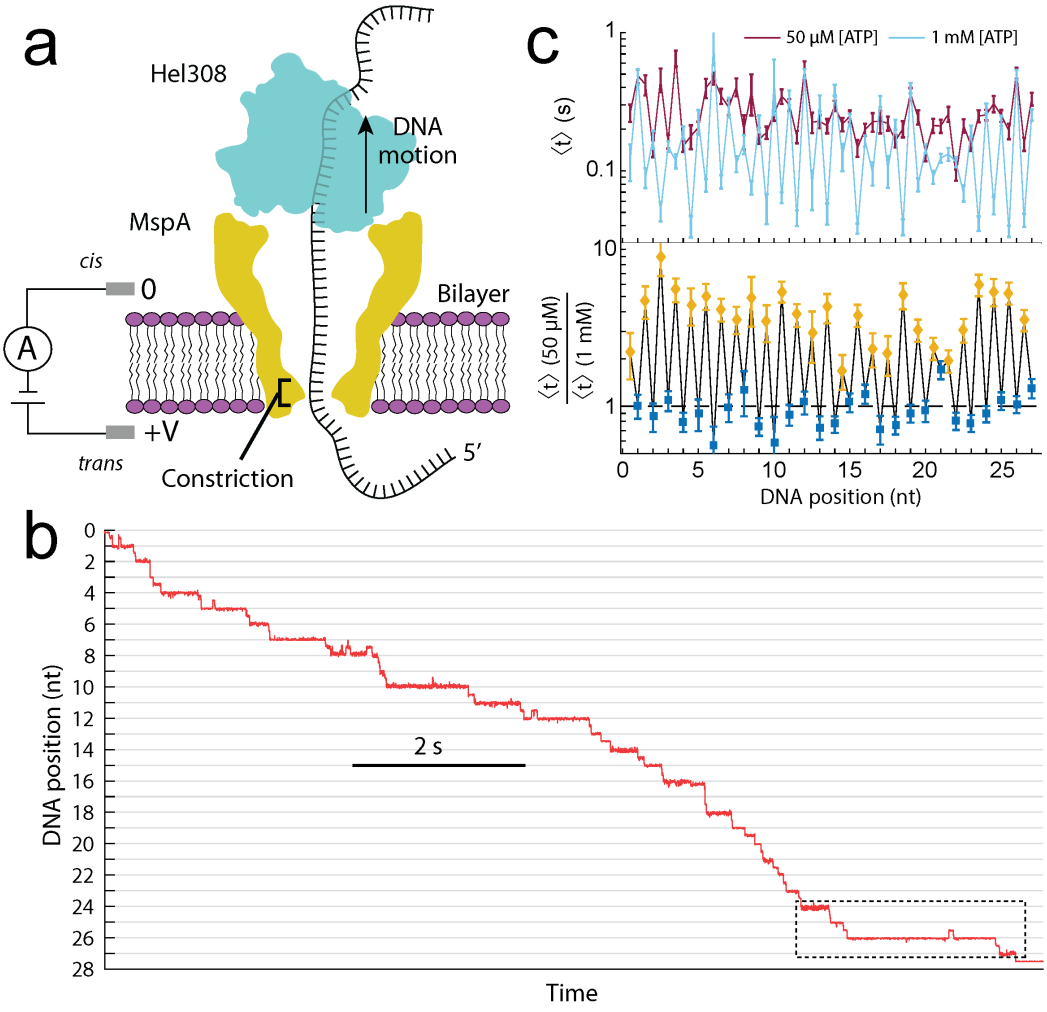

Figure 3.1: SPRNT on Hel308 Helicase

(a) A schematic of Hel308 translocase experiment. A single Hel308 molecule on an MspA pore draws the DNA out of the pore. The voltage applies a force to the DNA in the constriction, resulting in a force on the enzyme. (b) Position versus time trace for a single Hel308 molecule moving on ssDNA. Note that sub-nt steps are well resolved. The section of data in the dashed box will be examined in figure 3.2. This data is sampled at 500 Hz. (c) (top) The average dwell time of Hel308 enzyme states versus DNA position at [ATP] = 50 μM (dark red) and [ATP] = 1000 μM (light blue). (bottom) The ratio of the two curves from above. The dwell time changes with [ATP] at half-integer nucleotide positions, while the dwell time does not change with [ATP] at integer nucleotide positions.

3.2 Kinetic Methods

With SPRNT it is possible to know whether an enzyme moves forwards or backwards as it transitions between observable kinetic states. We call DNA movement from 3′ to 5′ in the pore a 'forwards step', and DNA movement from 5′ to 3′ a 'backwards step.' We analyze the properties of forwards and backwards steps as a function of experimental conditions and DNA position. In addition to forwards and backwards steps, we also consider the initial conditions of each enzyme step [83, 84]. As an example, figure 3.2a shows a hypothetical kinetic model. The different rows represent observable states (labeled observable states 1, 2, and 3). Transitions within the rows cannot be observed by enzyme progression along the DNA, but their existence can be inferred by analyzing dwell time distributions at each DNA position (Fig. B.3, Appendix B.3). We call transitions within rows 'hidden chemical transitions'. The two paths (red and blue Fig. 3.2a) both pass through observable state 2 and conclude in observable state 3 (Fig. 3.2b). However, because the two paths proceed through different hidden chemical transitions within observable state 2, the underlying dwell time distributions for observable state 2 (Fig. 3.2b, inset) are different, emphasizing the importance of careful step classification. In this article we use the following notation for transitions between observable states: f|f for a forwards step following forwards step, f|b for a forwards step following a backwards step, and b|f for a backwards step following a forwards step. Backwards steps following backwards steps (b|b) were rare, and we did not analyze them in detail. We study f|f, f|b and b|f steps for both of Hel308s observable states ([ATP]-dependent and [ATP]-independent) using the methods of single-molecule enzyme dynamics [43, 53, 85] to reveal kinetic pathways of Hel308 translocation. Examples of these step types are shown in figure 3.2c.

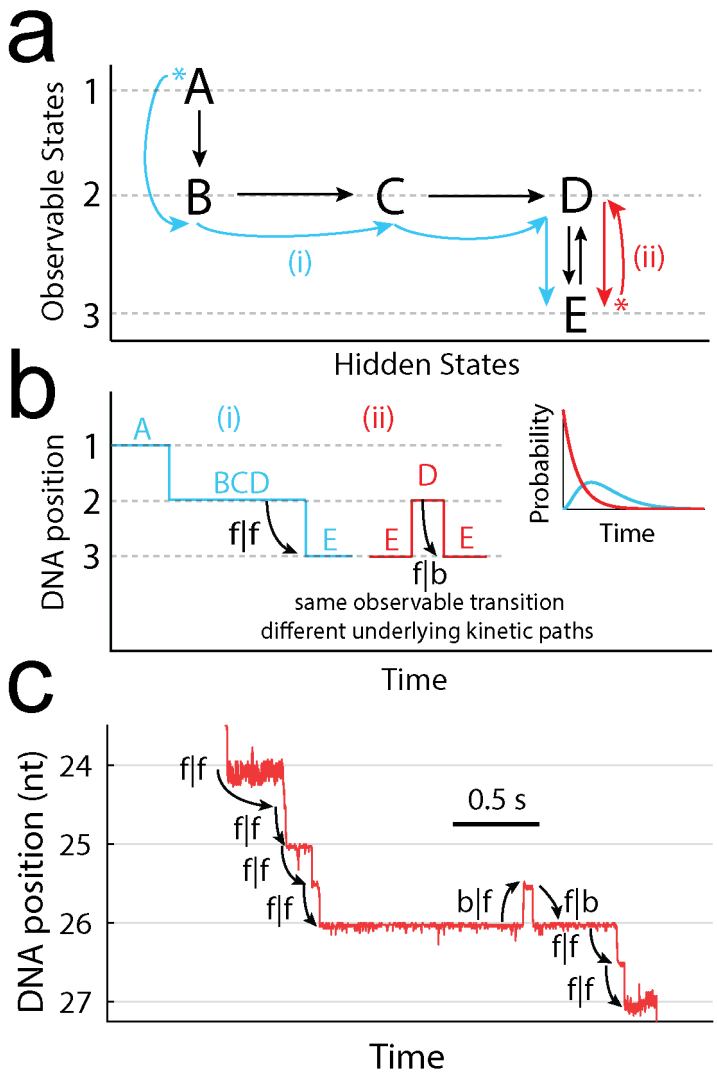

Figure 3.2: Methods for Analyzing Enzyme Kinetics with SPRNT

(a) A hypothetical kinetic model, with two possible enzyme paths through the model. In path i (blue), the progression of chemical states is $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$, resulting in three observed steps in SPRNT (1, 2 and 3). In path ii (red), the progression of chemical states is $E \rightarrow D \rightarrow E$, resulting again in three observed steps (3, 2 and 3). '*' indicates the initial kinetic state. (b) A hypothetical DNA positon vs time trace for the two paths shown in 3.2a. On average, the dwell time of the second step is longer in path (i) because that path goes through multiple hidden kinetic steps ($B \rightarrow C$, $C \rightarrow D$, and $D \rightarrow E$), whereas in path (ii) the dwell time is determined by $D \rightarrow E$ alone, resulting in a short average dwell time. (Inset) Hypothetical probability distribution of dwell times for path (i) and (ii) (blue, red respectively). Because path i progress through more hidden chemical transitions, the dwell time distribution is the convolution of several exponential processes. (c) Raw DNA position vs. time data trace for Hel308 on ssDNA (dashed box in Fig 3.1b), with step classifications indicated.

3.3 Results

3.3.1 ATP and ADP binding to Hel308

We varied [ATP] and [ADP] independently to analyze their effects on Hel308 translocation. Figure 3.3a shows the reaction rates for f|f [ATP]-dependent and f|f [ATP]-independent step types at several DNA positions as a function of [ATP] at [ADP] = 0. The data is binned by DNA position to probe for potential sequence-dependent kinetics. Unsurprisingly, rates of the individual f|f [ATP]-independent steps are unchanged over this [ATP] range. The rate of each f|f [ATP]-dependent step is well described by the Michaelis-Menten equation

$$rate \equiv \frac{1}{\langle t \rangle} = \frac{V \cdot [ATP]}{K + [ATP]}, \tag{3.1}$$

where V is the rate of the reaction at saturating [ATP], and K is the Michaelis constant. V and K are related to the underlying chemical rate constants. There is significant variation in the values of V and K at different DNA positions (Range: $V = 10 - 34\ s^{-1}$, $K = 66 - 412\ \mu M$, Fig. B.4,B.5, Table B.2, Appendix B.4), suggesting that the Hel308 stepping behavior depends on the DNA bases in Hel308. Because of the variation in the translocation rate with DNA position, we analyzed each DNA position individually to avoid averaging over sequence-dependent effects that may disguise the underlying kinetics of Hel308 translocation.

Figure 3.3b shows the average duration of f|f steps at varying [ADP] with [ATP] = 50 μM for the same DNA positions that are shown in figure 3.3a. The average duration of f|f [ATP]-dependent steps increases linearly with [ADP], implying that ADP acts as an inhibitor to the forwards progression Hel308 in the [ATP]-dependent step (Fig. B.6). The duration of f|f [ATP]-independent steps is unaffected by the presence of ADP, implying that ADP binding/unbinding to Hel308 does not occur during the [ATP]-independent step.

We analyzed b|f steps to understand backwards motion of Hel308. We first noticed that the probability of a b|f [ATP]-independent step changes with DNA position, ranging from 1% to 60% (Fig. 3.5, Appendix B.5), suggesting that the energetics landscape of Hel308 is modified by the DNA sequence. Figure 3.4a shows the probability of a b|f step at varying

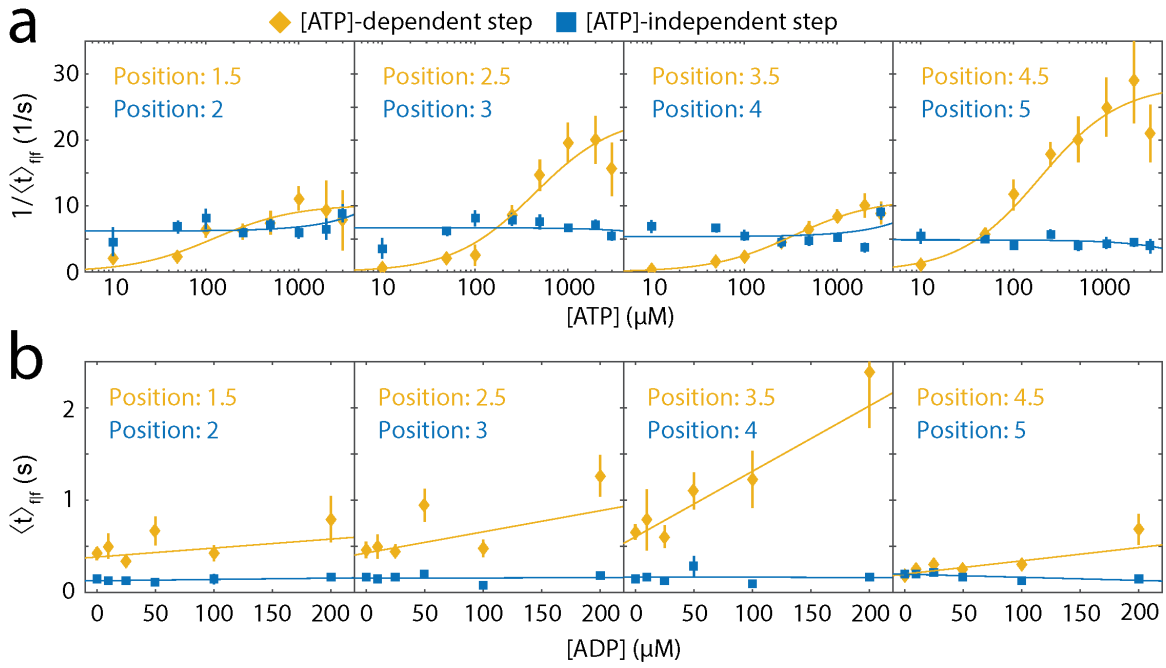

Figure 3.3: Analyzing ATP and ADP Dependence of Hel308 Forwards Steps

(a) The average rate of reaction for 4 sequential f|f [ATP]-dependent steps (yellow, half-integer positions) and 4 sequential f|f [ATP]-independent steps (blue, integer positions) as a function of [ATP] ([ADP] = 0). The DNA positions from figure 3.1b are displayed above. The x-axis is logarithmic. The yellow lines represent the best-fit Michaelis Menten equation to the [ATP]-dependent data. The blue lines are the weighted average of the [ATP]-independent data. (b) The mean dwell time for the same f|f [ATP]-dependent steps (yellow) and [ATP]-independent steps (blue) as in (a), as a function of the [ADP] ([ATP] = 50 μM). The DNA positions from figure 3.1b are displayed above. The yellow lines are the best linear fit to the data, while the blue lines are weighted average of the [ATP]-independent data. The x-axis is linear.

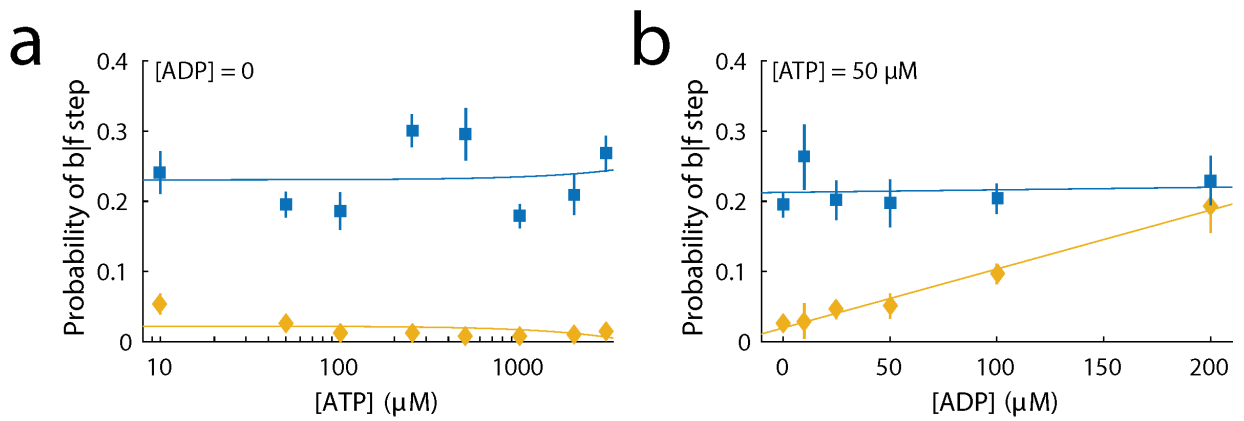

Figure 3.4: Analyzing ATP and ADP dependence of Hel308 Backwards Steps

(a) The probability of a b|f [ATP]-independent step (blue) and b|f [ATP]-dependent step (yellow) at varying [ATP] and fixed [ADP] = 0, averaged over all DNA positions. The weighted average of the [ATP]-independent data is plotted on top. The yellow line is fit based on our kinetic model for Hel308 (Fig. 3.10a, eq. B.45). The x-axis is logarithmic. (b) The probability of a b|f step, averaged over all DNA positions for the [ATP]-independent (blue) and [ATP]-dependent step (yellow) at varying [ADP] and fixed [ATP] = 50 μM. The blue line is the weighted average of the [ATP]-independent data. Error bars are S.E.M. The yellow line is fit based on our kinetic model for Hel308 (Fig. 3.10a, eq. B.45)

[ATP] ([ADP] = 0) for both [ATP]-dependent and [ATP]-independent step types, averaged over DNA position to accumulate sufficient statistics (N = 21 DNA positions). We find that the probability of a b|f [ATP]-independent step is independent of [ATP], while for the [ATP]-dependent step at very low ATP concentrations, the probability of a b|f step increases slightly. Figure 3.4b shows the probability of b|f step at varying [ADP] ([ATP] = 50 μM). The probability of a b|f [ATP]-independent step is independent of [ADP]. However, the probability of a b|f [ATP]-dependent step grows with [ADP] from ≈1% at [ADP] = 0 to ≈ 20% at [ADP] = 200 μM, demonstrating that ADP helps to drive Hel308 backwards along the DNA.

We compared the rate of f|f [ATP]-dependent steps to f|b [ATP]-dependent steps as a function of [ATP] for DNA positions in which the following [ATP]-independent step moved backwards often enough times to collect significant statistics (Fig. 3.6,B.7, Appendix B.6,B.7).

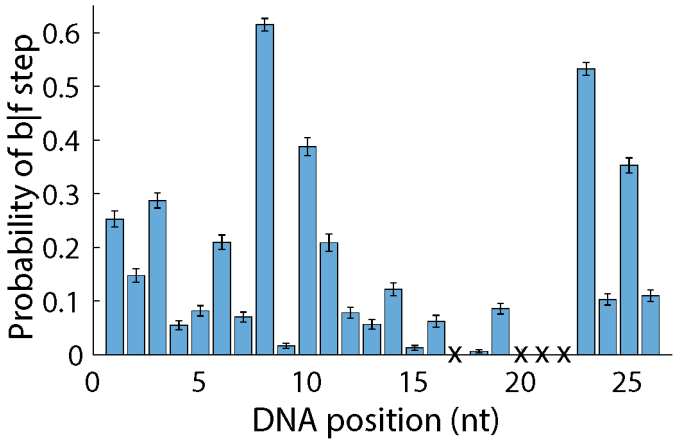

Figure 3.5: Probability of b|f [ATP]-independent Steps vs. DNA Position

Probability of a b|f step versus DNA position for the [ATP]-independent step. An 'x' indicates the the measurement could not be made due to adjacent ion current levels being too similar.

Interestingly, the kinetics of f|f and f|b [ATP]-dependent steps are nearly identical, with very similar values of V and K, which can be used to constrain the parameters of the underlying kinetics (Appendix B.6).

3.3.2 Analysis f|f, f|b and b|f [ATP]-independent dwell time distributions

Because the average dwell time of f|f [ATP]-independent steps is independent of both [ATP] and [ADP] (Fig. 3.3a,b), we assumed that the underlying dwell time distributions for each step were similarly unaffected by [ATP] and [ADP], enabling us to combine our data at various [ATP] and [ADP], yielding large statistics for each step. Figure 5 shows the dwell time distributions for f|f and f|b [ATP]-independent steps for each DNA position with N > 20 counts of the f|b step. We compared these distributions with several classes of exponential distribution function to probe the minimum number of hidden chemical transitions in the [ATP]-independent step (Fig. B.8, Table B.3, Appendix B.8). We found that the dwell time distributions of f|f [ATP]-independent steps are best described by the convolution of two exponential distributions, implying the existence of at least one hidden chemical step in the

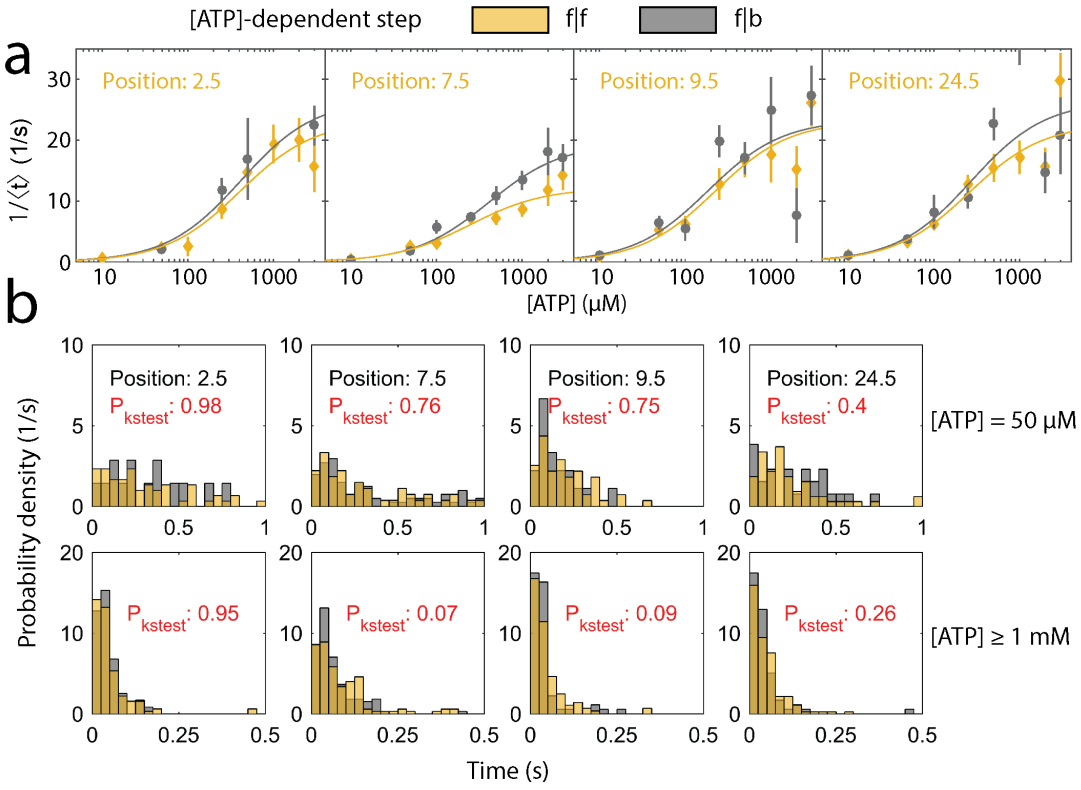

Figure 3.6: Comparing f|f and f|b [ATP]-dependent Steps

(a) The rate of reaction for f|f (yellow) and f|b (grey) [ATP]-dependent steps vs. [ATP] for several different DNA positions. The best fit to the Michaelis-Menten equation is plotted on top. The x-axis is logarithmic. (b) (Top) Dwell time distributions of the [ATP]-dependent step at given positions along the DNA at [ATP]=50 μM for f|f steps (yellow) and f|b steps (grey). The p-value for the two-sample KS test is displayed in red, indicating that the histograms are statistically indistinguishable.

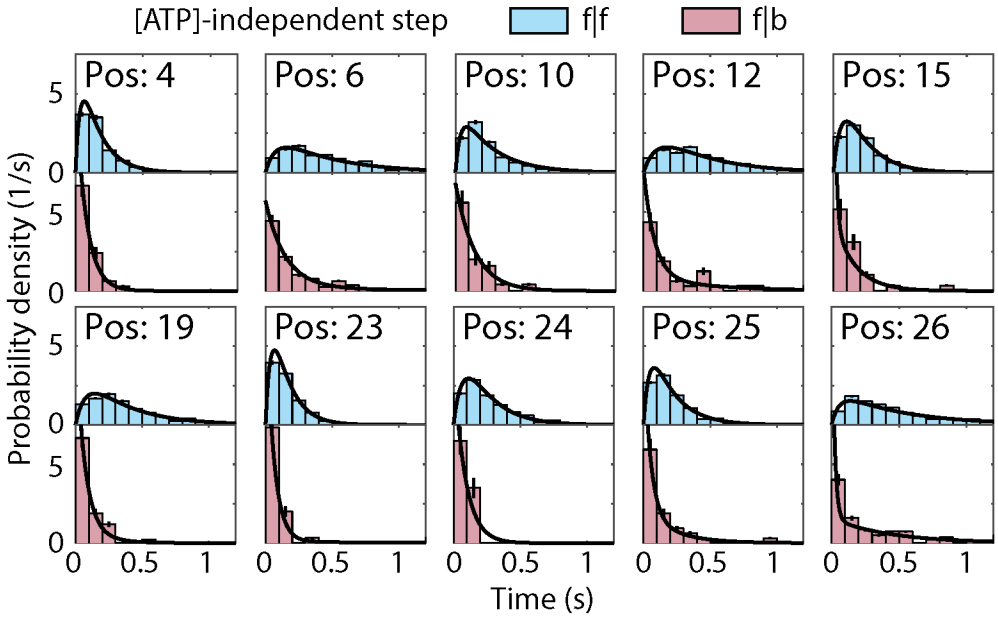

Figure 3.7: Comparing f|f and f|b [ATP]-independent Dwell Time Distributions

Probability distribution of dwell times for [ATP]-independent f|f steps (top, light blue) and f|b steps (bottom, light red). Histograms were constructed by collecting the measured dwell times for [ATP]-dependent steps at different DNA positions, averaging across all [ATP] and [ADP] conditions. The black lines drawn on the histograms are maximum likelihood estimates for two convolved exponential distributions (f|f steps eq. B.17) or a mixed two-exponential distribution (f|b steps, eq. B.19). Only DNA positions with $N > 20$ measurements of the f|b step are included. Error bars are S.E.M.

forwards progression of the [ATP]-independent step. In contrast, for f|b steps the dwell time distributions are best described by the sum of two exponentials. The different dwell time distributions for f|f and f|b [ATP]-independent steps indicates that the initial conditions of the [ATP]-independent step affect the reaction kinetics.

We similarly compared the dwell time distributions of f|f steps to b|f steps. Figure 3.8 shows the dwell time distributions for f|f and b|f steps for each DNA position with N > 50 counts of the b|f step. Surprisingly, we find that the f|f and b|f dwell time distributions are very similar [84], and for some DNA positions are statistically indistinguishable (8/16 DNA positions with p > 0.05). This suggests either that the rate-constants are finely tuned to

produce similar histograms, or that some population of b|f steps follow a chemical pathway parallel to f|f steps which instead results in an unproductive forwards step, resulting in a half-nucleotide backwards step. Interestingly, in each of the histograms in figure 3.8, the b|f step has a higher fraction of dwell times at low duration, suggesting that at least some population of the b|f steps are the reverse process of the previous f|f [ATP]-dependent step, or what we call an on-pathway backstep. Because we cannot determine whether these b|f steps are on-pathway or off-pathway, interpretation of the f|b [ATP]-dependent step data is more complicated, since we cannot know if the initial conditions of the [ATP]-dependent step are truly modified. This reasoning is further developed in Appendix B.6.

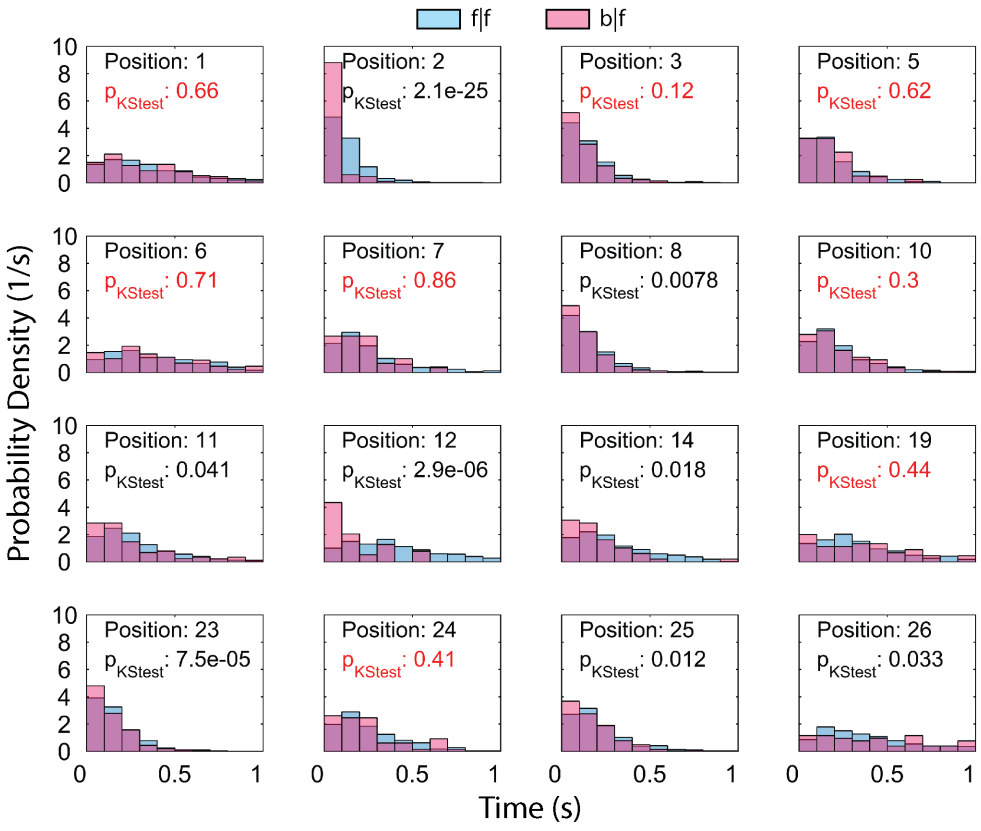

Figure 3.8: Comparing f|f and b|f [ATP]-independent Dwell Time Distributions

Dwell time distributions for f|f (blue) and b|f (pink) [ATP]-independent steps with $N \geq 50$ counts of the b|f step. The p-value for the two-sample KS test are displayed. In many instances the distributions are statistically indistinguishable (Those steps with $p > 0.05$, indicated in red).

3.3.3 Variation of voltage and temperature

SPRNT permits force spectroscopy by varying the voltage across the pore. Because force coupling in some mechano-chemical kinetic models depends on the substrate concentration [41], we calculated the mean dwell time of both f|f [ATP]-dependent and f|f [ATP]-independent step types at nearly-saturating ([ATP] = 500 μM) and sub-saturating ([ATP] = 50 μM) conditions at [ADP] = 0, averaged over DNA position (Fig. 3.9a,b, Table B.4, Appendix B.9). For both [ATP]-dependent and [ATP]-independent step types, we found that the average dwell time is independent of the applied voltage in the range 140 mV - 280 mV (force estimated as ≈30-65 pN [51]).

Changing the temperature of the solution can yield further insight into the energetics of helicase motion (Fig. 3.9c). We varied the temperature of the reaction volume from 22 °C to 45 °C while maintaining [ATP] = 500 μM and [ADP] = 0, and found that the average duration of both f|f [ATP]-dependent and f|f [ATP]-independent step types are well-described by an exponential function of the reciprocal temperature, consistent with Arrhenius equation (Appendix B.9). Because the kinetics are independent of the applied force, the activation energy is determined by fitting to Arrhenius equation. We find that $E_{[ATP]-dep} = 60 \pm 11\ kJ \cdot mol^{-1}$ and $E_{[ATP]-indep} = 77 \pm 15\ kJ \cdot mol^{-1}$. Because there are hidden chemical transitions in both the [ATP]-dependent and [ATP]-independent step types and because we averaged over DNA position, the calculated activation energies are the average activation energy for the rate-limiting substep of each observable step type.

3.3.4 Development of Hel308 kinetic model using the [ATP]-dependent step

We sought to construct the simplest possible kinetic model of Hel308 translocation consistent with each of the above observations. Because the [ATP]-independent step kinetics were independent of [ATP], [ADP], and applied force (Fig. 3.3,3.9), we focus on developing a model of the [ATP]-dependent step, and then hypothesize what processes occur during the [ATP]-independent step. We have established that ADP binding to Hel308 causes an

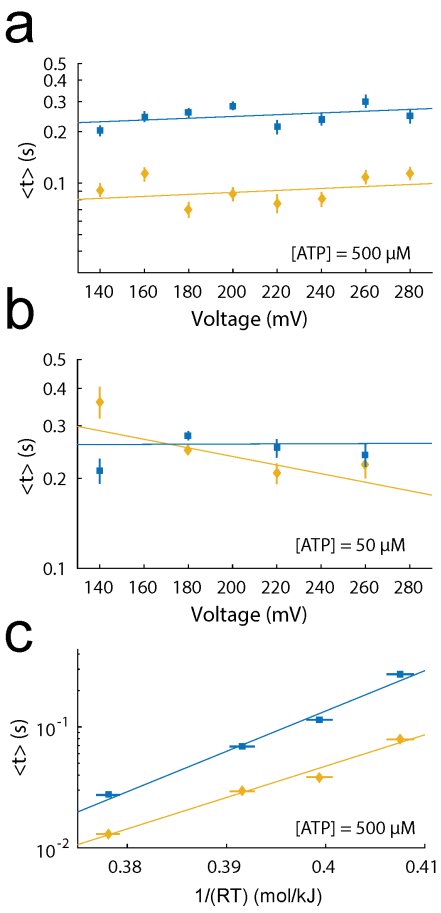

Figure 3.9: Effects of Varying Voltage and Temperature on Hel308 Kinetics

(a) The average dwell time of [ATP]-dependent (yellow) and [ATP]-independent (blue) steps averaged over DNA position vs. voltage at $[ATP] = 500\ \mu M$. The y-axis is logarithmic. Best fits to equation B.23 are plotted on top. The fit parameters are displayed in table B.4. (b) The average dwell time of [ATP]-dependent (yellow) and [ATP]-independent (blue) steps averaged over DNA position vs. voltage at $[ATP] = 50\ \mu M$. The y-axis is logarithmic. Best fits to equation B.23 are plotted on top. The fit parameters are displayed in table B.4. (c) The average dwell time taken over all [ATP]-dependent (yellow) and [ATP]-independent (blue) steps averaged over DNA position vs. inverse temperature at $[ATP] = 500\ \mu M$. The y-axis is logarithmic. Best fits to equation B.22 are plotted on top. All error bars are S.E.M. The error bars in temperature in (c) represent day-to-day and experiment-to-experiment fluctuations in temperature.

increase in the average duration of f|f [ATP]-dependent steps (Fig. 3.3b), implying that the ADP unbinding/binding is a hidden chemical transition within the [ATP]-dependent step. Similarly, we know that ATP binding must occur during the [ATP]-dependent step. Because the average duration of [ATP]-independent steps did not change when varying either [ATP] or [ADP], neither ATP nor ADP binding should occur in our model during the [ATP]-independent step.

Because ATP and ADP binding/unbinding occur during the [ATP]-dependent step, we are restricted to two general classes of model: those in which the ADP unbinding precedes the ATP binding, and vice versa. In Model 1 (Fig. 3.10a), ADP unbinds from Hel308 at the end of the previous hydrolysis cycle during the [ATP]-dependent step and is followed by ATP binding. Then the enzyme undergoes a conformational change to the [ATP]-independent step. Hel308 then proceeds through the remainder of the hydrolysis cycle in the [ATP]-independent step before transitioning back to the [ATP]-dependent step with ADP still bound. A similar model in which ATP directly induces the transition from the [ATP]-dependent step to the [ATP]-independent step is explored in Appendix B.12, and we find it to be inconsistent with the data. In Model 2 (Fig. 3.10b), the [ATP]-independent step concludes with a thermally driven conformational change that is rectified upon ATP binding during the [ATP]-dependent step. ATP is then hydrolyzed to ADP, whereupon ADP is released before Hel308 undergoes a conformational change back to the [ATP]-independent step.

We argue that the data is more consistent with Model 1 for two reasons:
1) Model 1 and Model 2 predict different dependences of the average dwell time of f|f [ATP]-dependent steps on [ATP] and [ADP] (Fig. B.9, Appendix B.10):

$$\langle t \rangle_{f|f,\ Model\ 1} = \frac{K + [ATP] + d \cdot [ADP]}{V \cdot [ATP]} \tag{3.2}$$

$$\langle t \rangle_{f|f,\ Model\ 1} = \frac{K' + [ATP] + d' \cdot [ADP] + e \cdot [ATP] \cdot [ADP]}{V' \cdot [ATP]} \tag{3.3}$$

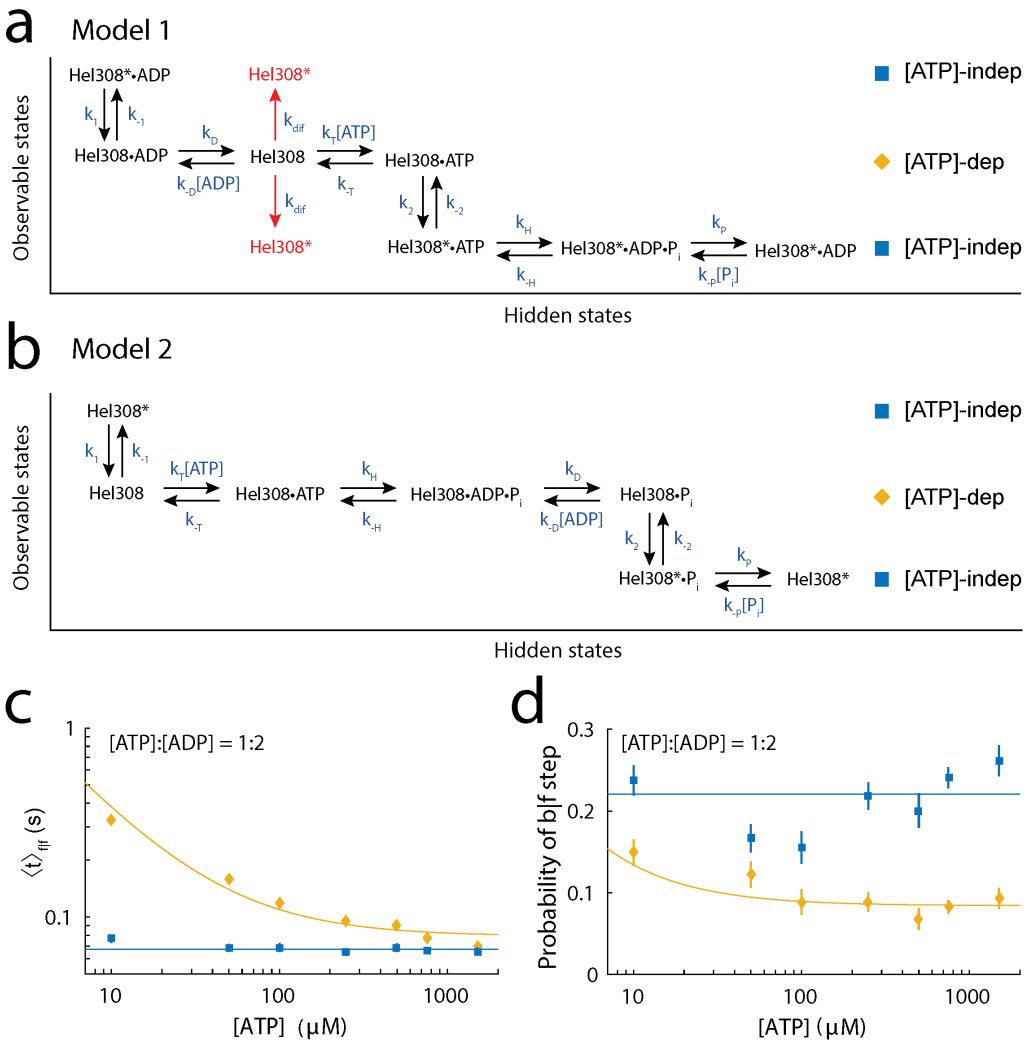
a
Model 1
Observable states
Hel308*•ADP
Hel308•ADP
Hel308*
Hel308
Hel308•ATP
Hel308*•ATP
Hel308*•ADP•Pi
Hel308*•ADP
[ATP]-indep
[ATP]-dep
[ATP]-indep
Hidden states
b
Model 2
Observable states
Hel308*
Hel308
Hel308•ATP
Hel308•ADP•Pi
Hel308•Pi
Hel308*•Pi
Hel308*
[ATP]-indep
[ATP]-dep
[ATP]-indep
Hidden states
c
[ATP]:[ADP] = 1:2
⟨t⟩flf (s)
1
0.1
10
100
1000
[ATP] (μM)
d
[ATP]:[ADP] = 1:2
Probability of bJf step
0.3
0.2
0.1
0
10
100
1000
[ATP] (μM)

Figure 3.10: Kinetic Model of Hel308 Translocation on ssDNA

(a) Kinetic Model 1. ADP bound to Hel308 in the [ATP]-dependent step at the end of the previous hydrolysis cycle is released. ATP then binds to Hel308, followed by a conformational change to the [ATP]-independent step. ATP is hydrolyzed to ADP, followed by relaxation to the [ATP]-dependent step. The placement of the phosphate unbinding step is done so as not to lead to any contradictions to the data. '*' indicates Hel308 is in the [ATP]-independent step. The red arrows indicate an alternative model, in which free Hel308 can diffuse between translocation states. (b) Kinetic Model 2. Hel308 undergoes a conformational change from the [ATP]-independent step to the [ATP]-dependent step which is rectified upon ATP binding. ATP is then hydrolyzed to ADP, which is released, followed by a relaxation to the [ATP]-independent step. The placement of the phosphate unbinding step is done somewhat arbitrarily, in order to maintain multiple rate constants for the [ATP]-independent step. (c) Average dwell time taken over DNA position of f|f [ATP]-dependent steps (yellow) and f|f [ATP]-independent steps (blue) at varied [ATP] while maintaining a constant ratio of [ATP]:[ADP] = 1:2. The best fit to equation 3.3 for the [ATP]-dependent step is plotted on top (yellow), yielding $e = 0$. The weighted average of [ATP]-independent step data is plotted on top (blue). (d) The probability of a b|f step averaged over DNA position for [ATP]-dependent (yellow) and [ATP]-independent (blue) steps at varied [ATP] while maintaining a constant ratio [ATP]:[ADP] = 1:2. The weighted average of the [ATP]-independent step data is plotted on top. The yellow line is fit based on Model 1 + diffusion (B.42). The experiments of (c) and (d) were done at elevated temperature T = 37 °C to increase the entry rate of DNA-Hel308 complexes into the pore. Error bars are S.E.M.

V and K are the standard Michaelis-Menten parameters. The $\prime$ in 3.3 indicates that the model parameters are related to the underlying rate constants differently in the two models (Appendix B.10). In Model 2, the term e couples the average dwell time of f|f [ATP]-dependent steps to the product [ATP][ADP]. Figure 3.10c shows the dwell time averaged over all DNA positions against the [ATP] at fixed [ATP]:[ADP] = 1:2 ([ATP]:[ADP] = 1:4, Fig. s14). Fitting to equation 3.3 for both experiments yields e $\approx$ 0, so that equation 3.3 takes the form of equation 3.2. In order for e to be 0 in Model 2, $k_H + k_{-H}$ must be vanishingly small (Appendix B.10), in support of Model 1.

2) In Model 1, ATP and ADP compete for the ATP binding site, implying that the probability of a b|f [ATP]-dependent step depends only on the ratio [ATP]:[ADP]. In Model 2, [ATP] and [ADP] contribute separately to the probability of a b|f [ATP]-dependent step (Fig. B.10, Appendix B.5, equations B.43-B.44). Figure 3.10d shows the probability of a b|f step for both observable step types as a function of [ATP] while maintaining a fixed ratio [ATP]:[ADP] = 1:2 ([ATP]:[ADP] = 1:4, Fig. B.11). We see that in each experiment the probability of a b|f step increases when [ATP] decreases, seemingly at odds with Model 1, and in support of Model 2. However, with a slight modification to Model 1, allowing for free diffusion of Hel308 with neither ATP nor ADP bound (Fig. 3.10a, red, one dimensional diffusion has been observed in RNA polymerase at low [NTP] [86]). By simultaneously fitting the probability of a b|f step against both [ATP] and [ADP] for Model 1, Model 2 and Model 1 + diffusion (Fig. B.12), we find that Model 1 + diffusion best describes the data. Given the previous data in support of Model 1, we argue that this simple addition to Model 1 resolves any tension between the data and the two models.

Assuming Model 1 (Fig. 3.10a) and analyzing the distribution of durations for each [ATP]-dependent f|f step, we evaluated the parameters $k_{\pm T}$, k_2, $k_{\pm D}$ for the [ATP]-dependent step at each DNA position (Fig. 3.10a, B.13,B.14, Table B.5,B.6, Appendix B.12). We find that on average $k_{-T} \approx 30\ s^{-1}$, $k_T \approx 0.3\ \mu M^{-1} \cdot s^{-1}$, $k_2 \approx 17\ s^{-1}$, $k_D \approx 170\ s^{-1}$ and k_{-D} 4 $\mu M^{-1} \cdot s^{-1}$. We can also use that $p_{b|f} \approx 0.01$ at [ADP] = 0 and saturating [ATP] to estimate that on average $k_{-1} \approx 2\ s^{-1}$. At saturating [ATP] the rate-limiting step of

f|f [ATP]-dependent steps is k_2, implying that the conformational change of Hel308 is the rate-limiting step.

3.4 Discussion

We used SPRNTs high spatio-temporal resolution to determine the mechanism of Hel308 translocation on ssDNA by analyzing previously unobservable transitions between two substates of its ATP hydrolysis cycle, while maintaining absolute registration of the position of Hel308 on the DNA to find evidence for sequence-dependent translocation of Hel308 on ssDNA. We analyzed forwards and backwards steps of both observable step types by varying [ATP], [ADP], force and temperature to provide insights into the translocation mechanism of the [ATP]-dependent step. By understanding the translocation mechanism of Hel308 on ssDNA, we may be able to understand the process by which Hel308 unwinds duplex DNA. By comparing measurements of Hel308 translocation on ssDNA with SPRNT measurements of Hel308 unwinding dsDNA, it should be possible to determine the chemical step at which duplex separation occurs, helping to determine whether Hel308 actively destabilizes the DNA duplex or whether Hel308 requires that the dsDNA dissociate thermally before stepping forward35.

For both [ATP]-dependent and [ATP]-independent step types, the dwell time distributions and probability of a backwards step are strongly dependent on DNA position. Because sub-nt steps are well-resolved and the DNA bases near the enzyme are being simultaneously sequenced by nanopore sequencing, SPRNT is uniquely suited among single-molecule techniques to examine DNA position-dependent kinetics. Sequence dependent unwinding kinetics have been observed in SFI/SFII helicase systems, due to the relative energies required to unwind GC versus AT base pairs [87, 88, 89]. However, sequence dependent translocation of a helicase on ssDNA, to the best of our knowledge, has not yet been observed. It is likely that DNA sequence in Hel308 affects the kinetics seen here. Because the translocation rate of Hel308 is independent of the applied force, alternative hypotheses such as sequence dependent force on the DNA in the nanopore39 are unlikely to cause the observed DNA position

dependent translocation. The crystal structure of a Hel308 helicase conjugated with DNA shows about 10 direct amino acid-DNA base interactions between the helicase and template DNA strand [60], suggesting that the role DNA sequence plays in determining translocation kinetic parameters may be complex. Further SPRNT studies to examine Hel308 translocation on either long DNA or carefully controlled short sequences will be required to fully evaluate DNA sequence-dependent phenomena. Other nucleic acid processing enzymes that have been measured with other single-molecule techniques may have similar sequence specific behavior that could not have been recognized and may be worth revisiting with SPRNT.

While SPRNT stems from nanopore DNA sequencing, the ability to perform in-depth kinetic analysis of enzymes using SPRNT can be used to improve nanopore sequencing. Nanopore sequencing is hindered, in part, by enzyme stepping behaviors such as backwards steps and missed ion current steps due to fast progression through the kinetic pathway. Mutations to promising enzymes can be investigated kinetically and selected for properties such as low probability of backstep, high throughput, and multiple rate constants per step (as opposed to a single rate-limiting step) to regularize the motion of the motor. An enzyme such as Hel308 with many kinetic substates that occur on similar time scales will have a more regular average duration per nucleotide when compared with an enzyme with a single dominant rate constant, which can help to determine the lengths of homopolymer sequences. Optimization of the controlling motor enzyme will be an important step in maximizing nanopore sequencing accuracy.

Chapter 4

INVESTIGATING THE EFFECTS OF DNA SEQUENCE ON HEL308 TRANSLOCATION ON SSDNA

We argued in the last chapter that the observed dependence of Hel308 kinetics on DNA position was due to the effect of the DNA sequence in Hel308. Here we investigate this claim experimentally. The work discussed in this chapter is ongoing, and the results are preliminary, but of significant interest to our work on Hel308 and using SPRNT to analyze sequence-dependent enzyme dynamics.

4.1 Introduction

DNA sequence is known to regulate enzyme dynamics in many enzyme systems such as DNA polymerases [90, 91], RNA polymerases [92], gene editing endonucleases such as CRISPR/Cas9 [93], and helicases through GC versus AT bond breaking [87, 88, 89]. To our knowledge, sequence dependent translocation of a helicase on ssDNA has yet to be observed. The goal of this chapter is to demonstrate that the kinetics of Hel308 translocation on ssDNA depends on the DNA sequence in Hel308 and to propose further experiments to investigate these phenomena.

SPRNT measures the position of DNA relative to the MspA constriction, but calculating the number of nucleotides between MspA and a relevant amino-acid residue is more complicated, especially in helicases. In RNA or DNA polymerase systems, one can starve the enzyme a single nucleotide triphosphate substrate. For example, in a DNAP experiment with substrate concentrations $[dATP] = [dCTP] = [dGTP] = 1\ mM$ and $[dTTP] = 10\ \mu M$ the dwell times of only those states in which a $dTTP$ is being incorporated into the DNAP will increase [31]. A cross correlation can then be used to determine the exact number of DNA

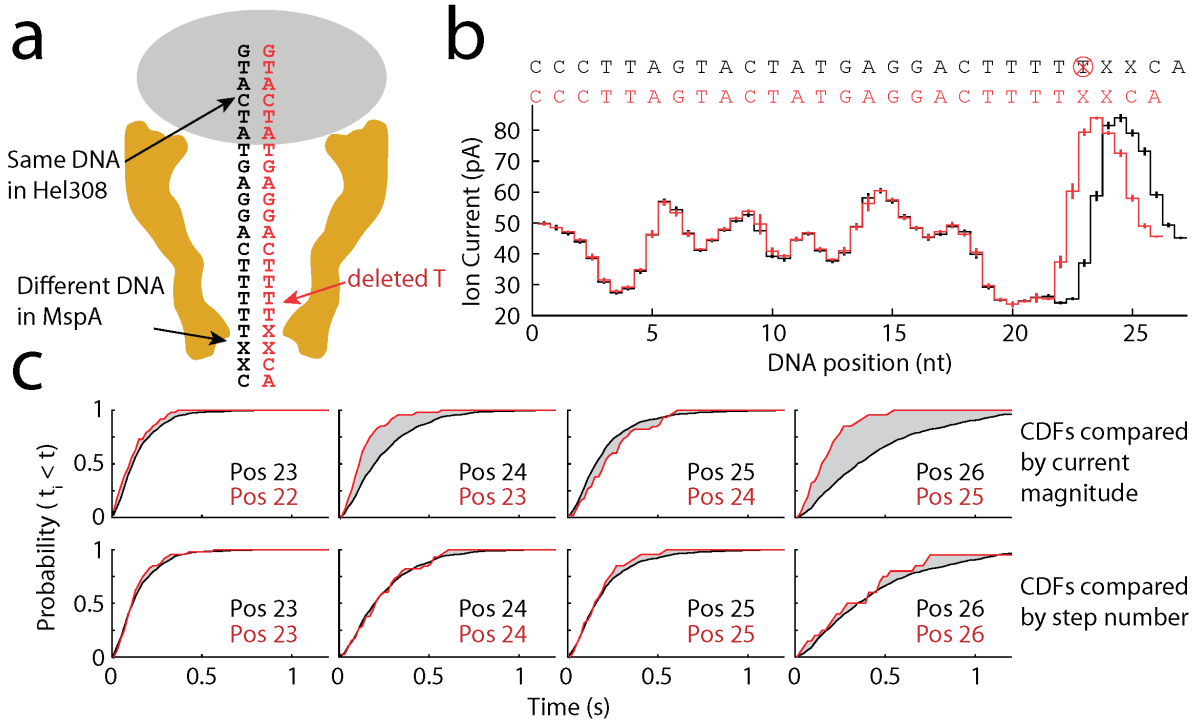

Figure 4.1: Analysis of Hel308 dwell-time distributions for two DNA Strands

(a) Experimental scheme. The red DNA strand is identical to the black, except for a single deleted T. (b) Ion current consensus plots for both DNA strands shown in (a). (c) Cumulative distribution functions for $f|f$ [ATP]-independent steps compared by current magnitude (top row) and by distance along Hel308 (bottom row).

bases between the MspA constriction and the enzyme active site. The same process does not work on helicases because each step uses the same fuel (ATP) to generate movement. Techniques such as calculating mutual information between dwell-time and DNA nucleotide in Hel308 can be powerful but require large amounts of data and sequence context that are not feasibly obtained at this time with SPRNT [94]. As such, we take the simplest possible approach in this chapter and examine Hel308 translocation on simple modifications of the DNA strand used in chapter 3 as well as translocation on homopolymeric strands.

4.2 Results

In chapter 3 we argued that because Hel308 kinetics were independent of the applied force, the nucleotides in the constriction could not affect the observed dwell-time distributions and therefore the observed change in kinetic parameters with DNA position must be due to the sequence in MspA. To test this hypothesis, we deleted a single thymine from the DNA sequence used in chapter 3, such that the same DNA sequence would be in Hel308, while causing a shift of the nucleotides within MspA's constriction (figure 4.1a,b). We then compared the following hypotheses:

- Hypothesis 1: The DNA nucleotides in the constriction cause the observed dwell-time distributions.

- Effect on dwell-time distributions 1: If we compare the dwell-time distributions by ion current they will match better than if matched by distance from Hel308.

- Hypothesis 2: The DNA nucleotides in Hel308 caused the observed dwell-time distributions.

- Effect on dwell-time distributions 2: The ion-current doesn't matter, only the distance from the enzyme affects the dwell-time distributions.

Figure 4.1c shows the empirical cumulative distribution function (CDF) of dwell times for each DNA sequence compared by ion current magnitude (top row) and by distance from Hel308 (bottom row). We find that the dwell-time distributions match much better when compared by distance from Hel308 ($p < 10^{-10}$, Appendix C.2), suggesting that hypothesis 2 better describes the data, and that the observed kinetics are generated by the DNA in Hel308. In addition, the probability of a $b|f$ step matches much better when aligned based on distance from Hel308 (data not shown).

To further investigate these effects we note that based on the crystal structure of Hel308 [60], Hel308 makes ≈ 10 base-specific contacts with the DNA (figure 4.2), suggesting that

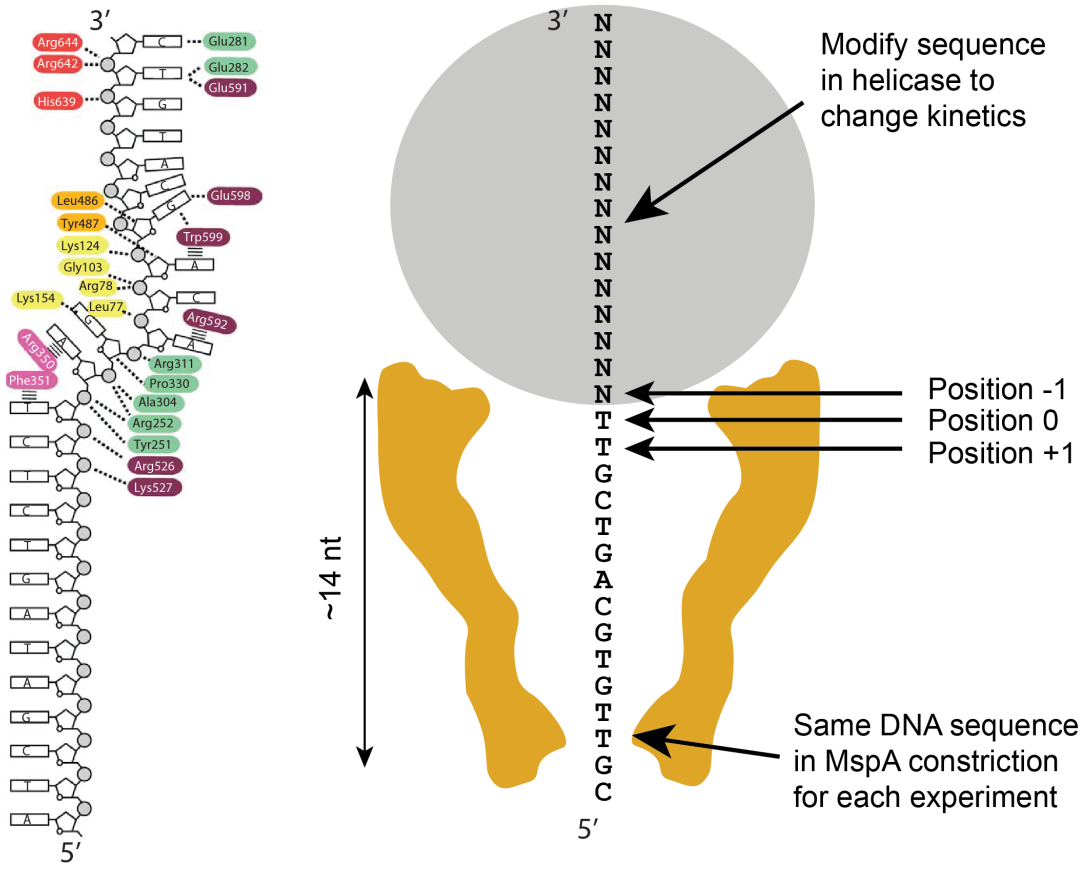

Figure 4.2: Experimental Scheme to Investigate Hel308 Sequence Dependence

(left) Adapted with permission from Macmillan Publishers Ltd: Nature Structural and Molecular Biology [60], copyright 2007. Interactions between Hel308 from *Archaeoglobus fulgidus* and the template DNA strand, determined from a crystal structure of Hel308 complexed with duplex DNA. Most of the DNA binding residues shown here are conserved in Hel308 from *Thermococcus gammatolerans*. Dotted lines indicate hydrogen bonds between amino acid residues and the DNA in Hel308. Bars indicate hydrophobic interactions between amino acid residues and DNA bases. (right) Experimental scheme. The sequence in Hel308 is modified (...NNN...), while the ion current of the downstream heteropolymer is measured by MspA. Illustration not to scale.

the relationship between DNA sequence and both dwell-times and probability of backwards steps may be highly complicated. To analyze this system, we use the experimental setup shown in figure 4.2, in which we use a DNA sequence in MspA that produces 'good' ion current profiles on which to do SPRNT calculations (i.e. heteropolymeric sequence that produces large changes in ion current), while varying the DNA bases in Hel308 to alter the dyanimcs of the helicase. We modify the 21 nt region upstream of the fixed sequence (DNA positions -1 to -21 in figure 4.2). As an initial test of sequence dependence, we analyzed Hel308 translocation on three separate DNA sequences in which positions -1 to -21 were set as homopolymers of A, C and T (hereafter: polyA, polyC and polyT, polyG forms complex secondary structures and therefore is not typically used). Figure 4.3a shows the ion current consensus for each strand. As expected, the ion current signatures are nearly identical for each of the DNA sequences from positions +2 onwards. Figures 4.3b-d show the average dwell time of the $f|f$ [ATP]-dependent step, $f|f$ [ATP]-independent step, and probability of a $b|f$ [ATP]-independent step, respectively, for each of the homopolymeric upstream sequences. Past position +18, each of the curves sync up, suggesting that the homopolymer is not affecting helicase dynamics past position +18. However, before position +18 the DNA sequence clearly modifies the kinetics. For example, in figure 4.3c, Hel308 tends to translocate fastest over polyT during the [ATP]-independent step. Interestingly, Hel308 backsteps often ($\approx$ %20 of steps) when walking over polyA and polyT, but almost never backsteps when walking over polyC (Fig. 4.3d).

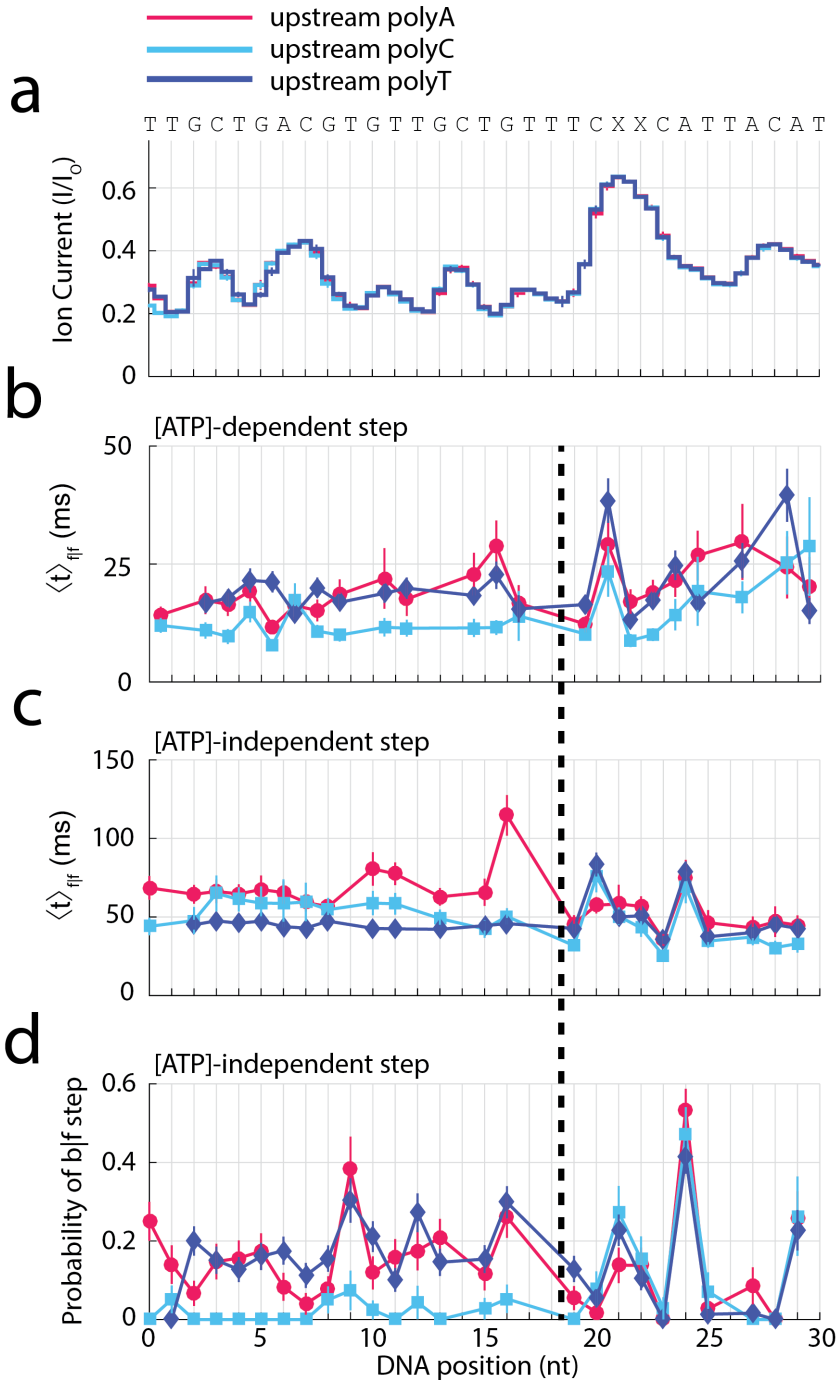
upstream polyA
upstream polyC
upstream polyT
a
T T G C T G A C G T G T T G C T G T T T C X X C A T T A C A T
Ion Current (I/I0)
0.6
0.4
0.2
0
b
[ATP]-dependent step
⟨t⟩fff (ms)
50
25
0
c
[ATP]-independent step
⟨t⟩fff (ms)
150
100
50
0
d
[ATP]-independent step
Probability of b|f step
0.6
0.4
0.2
0
0
5
10
15
20
25
30
DNA position (nt)

Figure 4.3: Hel308 Translocation on Homopolymer Sequences

(a) Consensus ion current sequences for three different DNA strands with the same heteropolymer as in figure 4.2b, but with different upstream sequence. As expected, the ion current is nearly identical for the three molecules. Colors correspond to the three strands: (pink) polyA (cyan) polyC (purple) polyT. The DNA sequence is written $3'$ to $5'$. (b) Average duration of $f|f$ [ATP]-dependent steps as a function of DNA position. Lines are guides to the eye. Gaps are positions where the ion current signal was too similar to another state to make reliable measurements of the average duration. (c) Average duration of $f|f$ [ATP]-independent steps as a function of DNA position. Lines are guides to the eye. Gaps are positions where the ion current signal was too similar to another state to make reliable measurements of the average duration. (d) Probability of a $b|f$ [ATP]-independent step as a function of DNA position. Lines are guides to the eye. Gaps are positions where the ion current signal was too similar to another state to make reliable measurement of $p_{b|f}$. The black dashed line indicates the approximate location past which the effects of the DNA sequence from positions -1 to -21 no longer has an effect.

For the rest of this chapter, we focus on analyzing how the probability of a $b|f$ [ATP]-independent step changes with DNA sequence, and hypothesize a mechanism for sequence specificity. We do not explore the [ATP]-dependent steps here because $b|f$ steps are sufficiently rare that there is not enough data to test sequence specificity. Our first goal was to determine which nucleotides in Hel308 affect the probability of a $b|f$ step. Figure 4.4a shows the probablity of a $b|f$ step for Hel308 translocating over three DNA sequences: polyC, a sequence of alternating AG (hereafter: upstreamAG), and a mixed sequence of polyC followed by repeating AGs (hereafter: upstreamMix). First we note that when translocating over upstreamAG that Hel308 has a much higher probability of backwards step at each DNA position compared to polyC. At even-numbered DNA positions the probability of a $b|f$ step is on average $\approx 10\%$, and at odd-numbered DNA positions is $\approx 50\%$. When Hel308 translocates over upstreamMix, the probability of a $b|f$ step is at first very small, matching the pattern of translocation over polyC from DNA positions 0 to +8. At position +9 the pattern abruptly shifts to match the pattern of translocation over upstreamAG. Because the transition in the upstreamMix sequence is so sharp, we conclude that only a small number of nucleotides determine the probability of a $b|f$ step; if more than several nucleotides were

determining the probability of a $b|f$ step then we would expect that the pattern of $b|f$ steps in the mixed sequence (blue) would have a region in which it did not match either polyC or upstreamAG. If we apply a shift of +20 nt to the DNA sequence, then the patterns of backsteps are matched up well with each of the three DNA sequences. After this shift has been applied, we can see that the probability of a $b|f$ step is largest when an adenine is present 20 nt upstream from MspA, and is smallest when cytosine is present 20 nt upstream from MspA.

Figure 4.4b shows the probability of a $b|f$ [ATP]-independent step for Hel308 translocation over polyC and a nearly identical sequence in which a single C was replaced with an A. Applying the +20 nt shift to the sequence as determined above, we see that the probability of a backstep increases only at the position of the substitution, consistent with the above observation that an adenine 20 nt upstream has an increased probability of a backwards step relative to cytosine. This suggests again that the number of nucleotides determining the probability of a $b|f$ step is small, and that the primary nucleotide determining [ATP]-independent kinetics is 20 nt upstream from the MspA constriction.

We sought to verify that the results of chapter 3 are consistent with the results discussed above. To that end, we took the 21-nt sequence we believed to cause the backstep pattern in chapter 3 (Figure 3.5) and placed it into the N_{21} section of the sequence used in figure 4.2(hereafter: port sequence). Figure 4.5a shows the probability of a $b|f$ [ATP]-independent step for the sequence shown in chapter 3 and the port sequence. Consistent with the results of figure 4.4, the probability of a $b|f$ step matches for the two sequences when the nucleotides 20 nt upstream from MspA match. After applying the +20 nt correction, we identified the nucleotides AGAC from positions +7 to +10 as those likely responsible for causing the large probability of a $b|f$ step at position +9. To test this hypothesis, we made sequential mutations to the DNA sequence (Fig. 4.5b): G→C at position +8 (cyan); G→C at +8 and A→C at +9 (blue); A→C at +7, G→C at +8 and A→C at +9 (maroon). Each of these mutations results in a decrease in the probability of a $b|f$ step at the position of the substitution, consistent with previous measurements. The mutation A→C at position +9

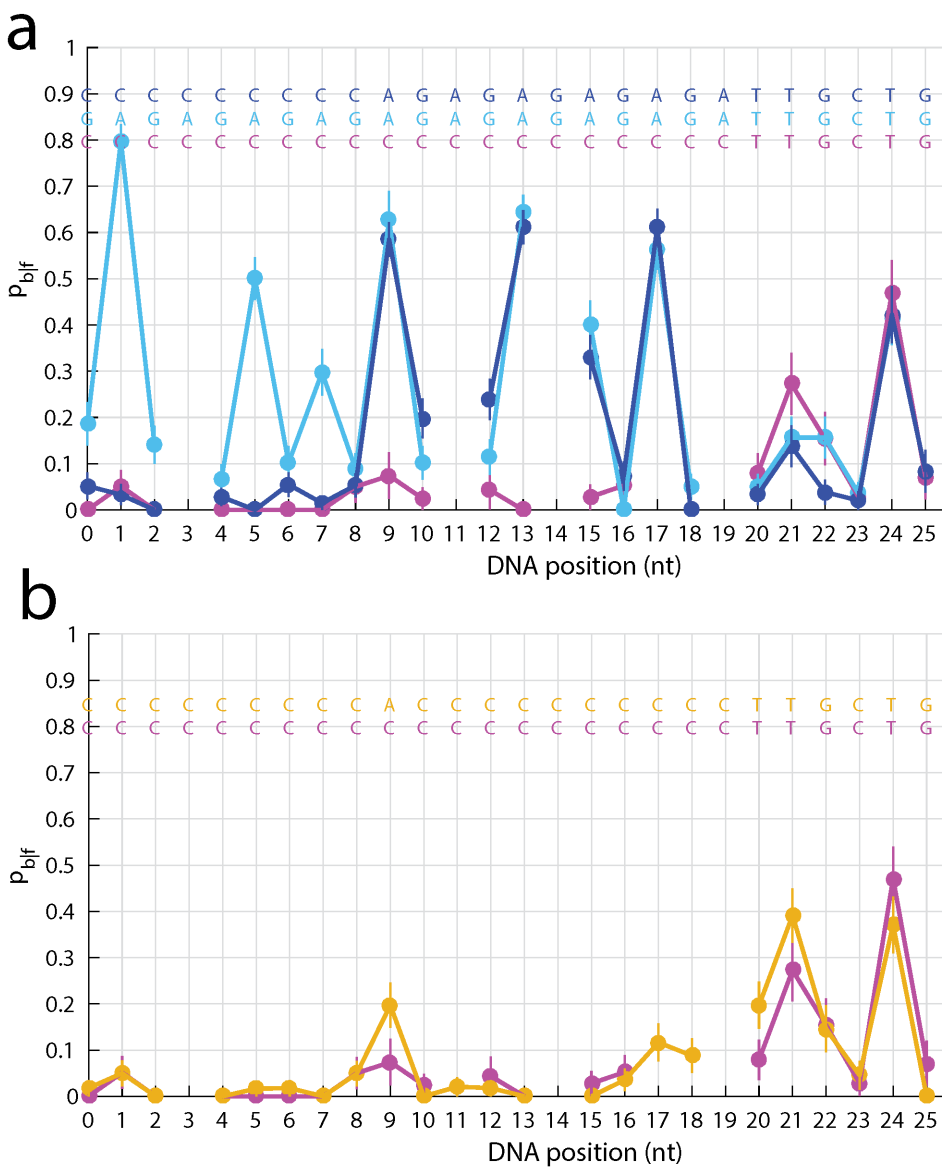

Figure 4.4: Determining sequence offset distance

Probability of a $b|f$ [ATP]-independent step for several DNA sequences. (a) polyC (magenta), upstreamAG (Cyan), upstreamMix (blue). (b) Mutating a single base in a cytosine background. (a) polyC (magenta), and polyC with a single Adenine substitution (yellow). Each DNA sequence was shifted by +20 nt to match the sequences with the underlying pattern. DNA sequences are written $3'$ to $5'$. Gaps indicate DNA positions where the probability could not be calculated because the ion current amplitudes were too similar.

results in a decrease in the probability of a $b|f$ step from $\approx 60\%$ to $\approx 5\%$. In fact, looking closely at figure 4.5 at positions 1, 3, 5, 10, 16, 18, and 23, whenever a Cytosine is present 20 nt upstream from the MspA constriction the probability of a $b|f$ step is very small.

4.3 Discussion

Based on the above results, and the Hel308 crystal structure ([60], Fig.4.2) we can the mechanism that leads to the sequence specificity in the [ATP]-independent step. One insight is provided by the low probability of a $b|f$ step whenever a cytosine is 20 nt upstream from the MspA constriction. Cytosine readily acts as both a hydrogen bond donor and acceptor, with three potential hydrogen bonds that can be made. Looking at the crystal structure (summarized in figure 4.2), the residue Glutamic acid (Glu) 598 can also form multiple hydrogen bonds as both a donor and acceptor, suggesting that it may form a highly stable complex with cytosine. Similarly, the high probability of a $b|f$ step tends to be associated with adenine, and may therefore be less stably bound to Glu598. In addition, the presence of the Tryptophan (Trp) 599 residue could participate in $\pi-\pi$ stacking interactions (electostatic quadrupole interactions) with adenine and guanine, promoting backwards motion of the DNA from Glu598 to Trp599. Since there are two amino-acid residues that appear to determine sequence specific translocation, the probability of a $b|f$ step will need to be measured for each three-nucleotide sequence to test this mechanism. In addition, the mutation of the Glu598 to Alanine would remove the hydrogen bonding interaction at that position, and will likely result in a higher probability of a $b|f$ step for each nucleotide, providing us with a rigorous test of the mechanism of sequence-specific translocation.

This analysis was done using only the probability of a backwards step. Additional analysis of the dwell-times of [ATP]-independent $f|f$ and $f|b$ steps, and their correlations with the probability of a $b|f$ step will yield further insights into the system. The results shown here demonstrate the ability of SPRNT to simultaneously use the DNA sequence information provided by nanopore sequencing and kinetics measurements to analyze the mechanism of sequence specificity in helicase systems.

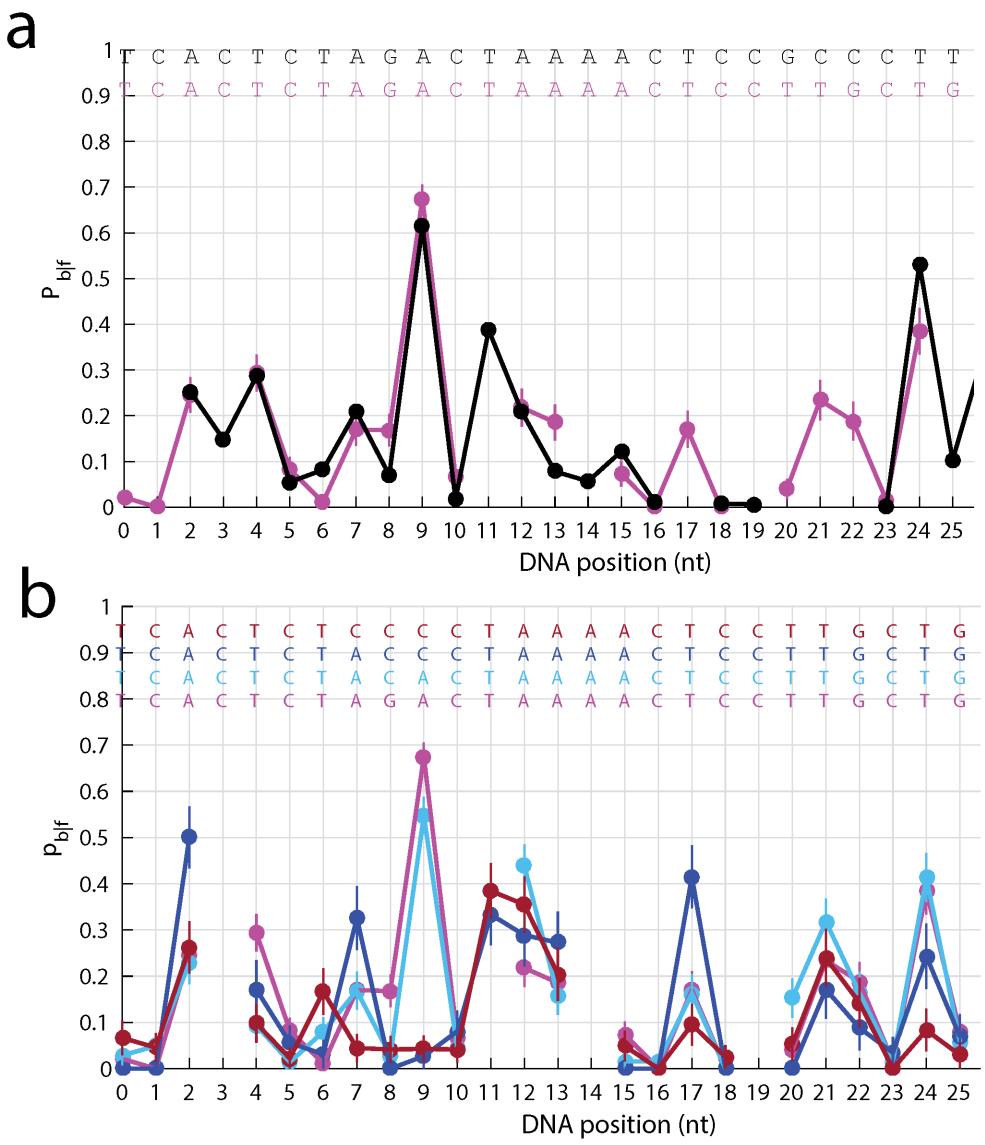

Figure 4.5: Transporting backsteps between sequences and mutating the DNA sequence

(a) probability of a $b|f$ [ATP]-independent step for the DNA sequence used in chapter 3 (black) and the port sequence (magenta). The DNA sequence has been offset by +20 nt to show the sequence in Hel308. (b) Making sequential mutations to the port sequence. The probability of a $b|f$ [ATP]-independent step for the port sequence (magenta), and mutated sequences G8C (cyan) G8C, A9C (blue), A7C, G8C, A9C (maroon). The DNA sequence has been offset by +20 nt to show the sequence in Hel308. The DNA sequences are written $3'$ to $5'$. Gaps indicate DNA positions where the probability could not be calculated because the ion current amplitudes were too similar.

Chapter 5

CONCLUSIONS

SPRNT is a new single-molecule enzyme probe that is being used to analyze enzymes at unprecedented spatiotemporal resolution, while also giving access to the underlying DNA sequence. In this thesis I've described the transition from MspA based nanopore DNA sequencing to SPRNT (chapter 2), and used SPRNT to analyze dynamics of the Hel308 helicase that can not be observed with any other single-molecule technique (chapter 3), including analysis of conditional dwell-time distributions, modeling of backwards enzyme steps, and dependence of both of these quantities on the DNA sequence (chapter 4). While great strides have been made by the UW nanopore group in the past several years developing SPRNT, there are several improvements that can be made to further improve SPRNT:

- The MspA nanopore can be inserted into the bilayer in a backwards orientation, which yields a different ion current to DNA base mapping, but may have other benefits, such as reducing the length of DNA between the enzyme and MspA's constriction, resulting in reduced fluctuations of the DNA within the pore. Similarly, DNA in the $3'$ orientation also has a different current-to-base map, meaning that there are four separate ion-current to DNA sequence maps which need to be measured to fully optimize SPRNT.

- While the force on the DNA has been estimated using DNA stretching curves, the result is model-dependent and has large errors ($\pm 10\ pN$). A direct force calibration with Optical Tweezers, Magnetic Tweezers, or AFM force spectroscopy in many different conditions (salt, temperature) is required to fully interpret SPRNT results. Because MspA pores are atomistically reproducible, the calibration needs to only be done once

in each condition and can then be used indefinitely.

- All of the experiments in this thesis were done using single-channel recording, however SPRNT will be much more useful once it has been fully parallelized, and SPRNT studies can be performed in a fraction of the time. Parallel platforms for nanopore DNA sequencing exist [95], however these platforms have not yet been expanded for use in SPRNT studies.

SPRNT results can be fed back into nanopore DNA sequencing to help improve base-calling algorithms. For example, if enzymes can be mutated to not backwards step, then it is less likely that a base will be called incorrectly. In addition, if an enzyme has multiple rate constants per nucleotide (e.g. Hel308), then the lengths of homopolymeric sequences can be determined more accurately by analyzing dwell-times. The development and feedback between nanopore DNA sequencing and SPRNT will lead to major improvements in both technologies.

BIBLIOGRAPHY

[1] Wolfram Saenger. Polymorphism of dna versus structural conservatism of rna: Classification of a-, b-, and z-type double helices. In *Principles of Nucleic Acid Structure*, pages 220–241. Springer, 1984.

[2] Ligang Wu and Joel G Belasco. Let me count the ways: mechanisms of gene regulation by mirnas and sirnas. *Molecular cell*, 29(1):1–7, 2008.

[3] Cecilia Guerrier-Takada, Katheleen Gardiner, Terry Marsh, Norman Pace, and Sidney Altman. The rna moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

[4] Kelly Kruger, Paula J Grabowski, Arthur J Zaug, Julie Sands, Daniel E Gottschling, and Thomas R Cech. Self-splicing rna: autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. *cell*, 31(1):147–157, 1982.

[5] Michael P Robertson and Gerald F Joyce. The origins of the rna world. *Cold Spring Harbor perspectives in biology*, 4(5):a003608, 2012.

[6] C NICK Pace, BRET A Shirley, M McNutt, and K Gajiwala. Forces contributing to the conformational stability of proteins. *The FASEB journal*, 10(1):75–83, 1996.

[7] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[8] JR Zacharias, S Millman, P Kusch, et al. A new method of measuring nuclear magnetic moment. *Physical Review*, 53:318–318, 1938.

[9] Mitch McVey, Varandt Y Khodaverdian, Damon Meyer, Paula Gonçalves Cerqueira, and Wolf-Dietrich Heyer. Eukaryotic dna polymerases in homologous recombination. *Annual Review of Genetics*, 50:393–421, 2016.

[10] P Cramer, K-J Armache, S Baumli, S Benkert, F Brueckner, C Buchen, GE Damsma, S Dengl, SR Geiger, AJ Jasiak, et al. Structure of eukaryotic rna polymerases. *Annu. Rev. Biophys.*, 37:337–352, 2008.

[11] Margaret E Fairman-Williams, Ulf-Peter Guenther, and Eckhard Jankowsky. Sf1 and sf2 helicases: family matters. *Current opinion in structural biology*, 20(3):313–324, 2010.

[12] Manfred Schliwa and Günther Woehlke. Molecular motors. *Nature*, 422(6933):759–765, 2003.

[13] Joseph L Kim, Kurt A Morgenstern, James P Griffith, Maureen D Dwyer, John A Thomson, Mark A Murcko, Chao Lin, and Paul R Caron. Hepatitis c virus ns3 rna helicase domain with a bound oligonucleotide: the crystal structure provides insights into the mode of unwinding. *Structure*, 6(1):89–100, 1998.

[14] Roger H Miller and Robert H Purcell. Hepatitis c virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups. *Proceedings of the National Academy of Sciences*, 87(6):2057–2061, 1990.

[15] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[16] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[17] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[18] Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10):1146–1153, 2008.

[19] Edward Abrahams and Mike Silver. The case for personalized medicine. *Journal of Diabetes Science and Technology*, 2009.

[20] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[21] Cees Dekker. Solid-state nanopores. *Nature nanotechnology*, 2(4):209–215, 2007.

[22] Langzhou Song, Michael R Hobaugh, Christopher Shustak, Stephen Cheley, Hagan Bayley, and J Eric Gouaux. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science*, 274(5294):1859–1865, 1996.

[23] Wenonah Vercoutere, Stephen Winters-Hilt, Hugh Olsen, David Deamer, David Haussler, and Mark Akeson. Rapid discrimination among individual dna hairpin molecules at single-nucleotide resolution using an ion channel. *Nature biotechnology*, 19(3):248–252, 2001.

[24] Qingtao Li, Qing Zhao, Bo Lu, Hengbin Zhang, Song Liu, Zhipeng Tang, Lijia Qu, Rui Zhu, Jingmin Zhang, Liping You, et al. Size evolution and surface characterization of solid-state nanopores in different aqueous solutions. *Nanoscale*, 4(5):1572–1576, 2012.

[25] Eric Beamish, Harold Kwok, Vincent Tabard-Cossa, and Michel Godin. Fine-tuning the size and minimizing the noise of solid-state nanopores. *JoVE (Journal of Visualized Experiments)*, (80):e51081–e51081, 2013.

[26] Tom Z Butler, Mikhail Pavlenok, Ian M Derrington, Michael Niederweis, and Jens H Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Sciences*, 105(52):20647–20652, 2008.

[27] Ian M Derrington, Tom Z Butler, Marcus D Collins, Elizabeth Manrao, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Nanopore dna sequencing with mspa. *Proceedings of the National Academy of Sciences*, 107(37):16060–16065, 2010.

[28] Michael Faller, Michael Niederweis, and Georg E Schulz. The structure of a mycobacterial outer-membrane channel. *Science*, 303(5661):1189–1192, 2004.

[29] Elizabeth A Manrao, Ian M Derrington, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Nucleotide discrimination with dna immobilized in the mspa nanopore. *PloS one*, 6(10):e25723, 2011.

[30] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of dna in a nanopore at 5-a precision. *Nature biotechnology*, 30(4):344–348, 2012.

[31] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading dna at single-nucleotide resolution with a mutant mspa nanopore and phi29 dna polymerase. *Nature biotechnology*, 30(4):349–353, 2012.

[32] Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Andrew Adey, Ian C Nova, Jonathan M Craig, Kyle W Langford, Jenny Mae Samson, Riza Daza, et al. Decoding long nanopore sequencing reads of natural dna. *Nature biotechnology*, 32(8):829–833, 2014.

[33] Andrew H Laszlo, Ian M Derrington, Henry Brinkerhoff, Kyle W Langford, Ian C Nova, Jenny Mae Samson, Joshua J Bartlett, Mikhail Pavlenok, and Jens H Gundlach. De-

tection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore mspa. *Proceedings of the National Academy of Sciences*, 110(47):18904–18909, 2013.

[34] Jacob Schreiber, Zachary L Wescoe, Robin Abu-Shumays, John T Vivian, Baldandorj Baatar, Kevin Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual dna strands. *Proceedings of the National Academy of Sciences*, 110(47):18910–18915, 2013.

[35] Zachary L Wescoe, Jacob Schreiber, and Mark Akeson. Nanopores discriminate among five c5-cytosine variants in dna. *Journal of the American Chemical Society*, 136(47):16582–16587, 2014.

[36] Zachary A Lewis, Shinji Honda, Tamir K Khlafallah, Jennifer K Jeffress, Michael Freitag, Fabio Mohn, Dirk Schübeler, and Eric U Selker. Relics of repeat-induced point mutation direct heterochromatin formation in neurospora crassa. *Genome research*, 19(3):427–437, 2009.

[37] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. Minion nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*, 33(3):296–300, 2015.

[38] George Edward Briggs and John Burdon Sanderson Haldane. A note on the kinetics of enzyme action. *Biochemical Journal*, pages 338–339, 1925.

[39] Leonor Michaelis and Maud L Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[40] ME Stroppolo, M Falconi, AM Caccuri, and A Desideri. Superefficient enzymes. *Cellular and Molecular Life Sciences*, 58(10):1451, 2001.

[41] David Keller and Carlos Bustamante. The mechanochemistry of molecular motors. *Biophysical Journal*, 78(2):541–556, 2000.

[42] H Peter Lu, Luying Xun, and X Sunney Xie. Single-molecule enzymatic dynamics. *Science*, 282(5395):1877–1882, 1998.

[43] Hong Qian and Elliot L Elson. Single-molecule enzymology: stochastic michaelis–menten kinetics. *Biophysical chemistry*, 101:565–576, 2002.

[44] Jeffrey R Moffitt, Yann R Chemla, Steven B Smith, and Carlos Bustamante. Recent advances in optical tweezers. *Annu. Rev. Biochem.*, 77:205–228, 2008.

[45] David Dulin, Tao Ju Cui, Jelmer Cnossen, Margreet W Docter, Jan Lipfert, and Nynke H Dekker. High spatiotemporal-resolution magnetic tweezers: Calibration and applications for dna dynamics. *Biophysical journal*, 109(10):2113–2125, 2015.

[46] Hajin Kim and Taekjip Ha. Single-molecule nanometry for biological physics. *Reports on Progress in Physics*, 76(1):016601, 2012.

[47] Elio A Abbondanzieri, William J Greenleaf, Joshua W Shaevitz, Robert Landick, and Steven M Block. Direct observation of base-pair stepping by rna polymerase. *Nature*, 438(7067):460–465, 2005.

[48] Timothée Lionnet, Michelle M Spiering, Stephen J Benkovic, David Bensimon, and Vincent Croquette. Real-time observation of bacteriophage t4 gp41 helicase reveals an unwinding mechanism. *Proceedings of the National Academy of Sciences*, 104(50):19790–19795, 2007.

[49] Taekjip Ha, Alice Y Ting, Joy Liang, W Brett Caldwell, Ashok A Deniz, Daniel S Chemla, Peter G Schultz, and Shimon Weiss. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proceedings of the National Academy of Sciences*, 96(3):893–898, 1999.

[50] Andrew H Laszlo, Ian M Derrington, and Jens H Gundlach. Mspa nanopore as a single-molecule tool: From sequencing to sprnt. *Methods*, 2016.

[51] Ian M Derrington, Jonathan M Craig, Eric Stava, Andrew H Laszlo, Brian C Ross, Henry Brinkerhoff, Ian C Nova, Kenji Doering, Benjamin I Tickman, Mostafa Ronaghi, et al. Subangstrom single-molecule measurements of motor proteins using a nanopore. *Nature biotechnology*, 33(10):1073–1075, 2015.

[52] Taekjip Ha, Alexander G Kozlov, and Timothy M Lohman. Single molecule views of protein movement on single stranded dna. *Annual review of biophysics*, 41:295, 2012.

[53] Gregory K Schenter, H Peter Lu, and X Sunney Xie. Statistical analyses and theoretical models of single-molecule enzymatic dynamics. *The Journal of Physical Chemistry A*, 103(49):10477–10488, 1999.

[54] Steven B Smith, Yujia Cui, and Carlos Bustamente. Overstretching b-dna: the elastic response of individual double-stranded and single-stranded dna molecules. *Science*, 271(5250):795, 1996.

[55] Alessandro Bosco, Joan Camunas-Soler, and Felix Ritort. Elastic properties and secondary structure formation of single-stranded dna at monovalent and divalent salt conditions. *Nucleic acids research*, 42(3):2064–2074, 2014.

[56] Colin P Guy and Edward L Bolt. Archaeal hel308 helicase targets replication forks in vivo and in vitro and unwinds lagging strands. *Nucleic acids research*, 33(11):3678–3690, 2005.

[57] Isabel L Woodman, Geoffrey S Briggs, and Edward L Bolt. Archaeal hel308 domain v couples dna binding to atp hydrolysis and positions dna for unwinding over the helicase ratchet. *Journal of molecular biology*, 374(5):1139–1144, 2007.

[58] Isabel L Woodman and Edward L Bolt. Winged helix domains with unknown function in hel308 and related helicases. *Biochemical Society Transactions*, 39(1):140–144, 2011.

[59] Isabel L Woodman and Edward L Bolt. Molecular biology of hel308 helicase in archaea. *Biochemical Society Transactions*, 37(1):74–78, 2009.

[60] Katharina Büttner, Sebastian Nehring, and Karl-Peter Hopfner. Structural basis for dna duplex separation by a superfamily-2 helicase. *Nature structural & molecular biology*, 14(7):647–652, 2007.

[61] Sua Myong, Michael M Bruno, Anna M Pyle, and Taekjip Ha. Spring-loaded mechanism of dna unwinding by hepatitis c virus ns3 helicase. *Science*, 317(5837):513–516, 2007.

[62] Kristina M Herbert, William J Greenleaf, and Steven M Block. Single-molecule studies of rna polymerase: motoring along. *Annual review of biochemistry*, 77:149, 2008.

[63] Hong Yin, Michelle D Wang, Karel Svoboda, Robert Landick, et al. Transcription against an applied force. *Science*, 270(5242):1653, 1995.

[64] Michelle D Wang, Mark J Schnitzer, Hong Yin, Robert Landick, Jeff Gelles, and Steven M Block. Force and velocity measured for single molecules of rna polymerase. *Science*, 282(5390):902–907, 1998.

[65] Krzysztof Sozański, Felix Ruhnow, Agnieszka Wiśniewska, Marcin Tabaka, Stefan Diez, and Robert Hołyst. Small crowders slow down kinesin-1 stepping by hindering motor domain diffusion. *Physical review letters*, 115(21):218102, 2015.

[66] Koen Visscher, Mark J Schnitzer, and Steven M Block. Single kinesin molecules studied with a molecular force clamp. *Nature*, 400(6740):184–189, 1999.

[67] Yannick Rondelez, Guillaume Tresset, Takako Nakashima, Yasuyuki Kato-Yamada, Hiroyuki Fujita, Shoji Takeuchi, and Hiroyuki Noji. Highly coupled atp synthesis by f1-atpase single molecules. *Nature*, 433(7027):773–777, 2005.

[68] Robert M Brosh Jr. Dna helicases involved in dna repair and their roles in cancer. *Nature Reviews Cancer*, 13(8):542–558, 2013.

[69] Li Fan, Jill O Fuss, Quen J Cheng, Andrew S Arvai, Michal Hammel, Victoria A Roberts, Priscilla K Cooper, and John A Tainer. Xpd helicase structures and activities: insights into the cancer and aging phenotypes from xpd mutations. *Cell*, 133(5):789–800, 2008.

[70] Marise R Heerma van Voss, Farhad Vesuna, Kari Trumpi, Justin Brilliant, Liudmila L Kodach, Folkert HM Morsink, G Johan A Offerhaus, Horst Buerger, Elsken van der Wall, Paul J van Diest, et al. Identification of the dead box rna helicase ddx3 as a therapeutic target in colorectal cancer, 2015.

[71] Payam Mohaghegh, Julia K Karow, Robert M Brosh Jr, Vilhelm A Bohr, and Ian D Hickson. The blooms and werners syndrome proteins are dna structure-specific helicases. *Nucleic acids research*, 29(13):2843–2849, 2001.

[72] Alicia K Byrd and Kevin D Raney. Superfamily 2 helicases. *Frontiers in bioscience (Landmark edition)*, 17:2070, 2012.

[73] Timothy M Lohman, Eric J Tomko, and Colin G Wu. Non-hexameric dna helicases and translocases: mechanisms and regulation. *Nature Reviews Molecular Cell Biology*, 9(5):391–401, 2008.

[74] Agnieszka A Tafel, Leonard Wu, and Peter J McHugh. Human hel308 localizes to damaged replication forks and unwinds lagging strand structures. *Journal of Biological Chemistry*, 286(18):15832–15840, 2011.

[75] Taekjip Ha, Thilo Enderle, DF Ogletree, Daniel S Chemla, Paul R Selvin, and Shimon Weiss. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences*, 93(13):6264–6268, 1996.

[76] Matthew J Comstock, Kevin D Whitley, Haifeng Jia, Joshua Sokoloski, Timothy M

Lohman, Taekjip Ha, and Yann R Chemla. Direct observation of structure-function relationship in a nucleic acid–processing enzyme. *Science*, 348(6232):352–354, 2015.

[77] Jae Young Lee and Wei Yang. Uvrd helicase unwinds dna one base pair at a time by a two-part power stroke. *Cell*, 127(7):1349–1360, 2006.

[78] Jeehae Park, Sua Myong, Anita Niedziela-Majka, Kyung Suk Lee, Jin Yu, Timothy M Lohman, and Taekjip Ha. Pcra helicase dismantles reca filaments by reeling in dna in uniform steps. *Cell*, 142(4):544–555, 2010.

[79] Sophie Dumont, Wei Cheng, Victor Serebrov, Rudolf K Beran, Ignacio Tinoco, Anna Marie Pyle, and Carlos Bustamante. Rna translocation and unwinding mechanism of hcv ns3 helicase and its coordination by atp. *Nature*, 439(7072):105–108, 2006.

[80] Gábor M Harami, Yeonee Seol, Junghoon In, Veronika Ferencziová, Máté Martina, Máté Gyimesi, Kata Sarlós, Zoltán J Kovács, Nikolett T Nagy, Yuze Sun, et al. Shuttling along dna and directed processing of d-loops by recq helicase support quality control of homologous recombination. *Proceedings of the National Academy of Sciences*, page 201615439, 2017.

[81] Maria Spies. Two steps forward, one step back: determining xpd helicase mechanism by single-molecule fluorescence and high-resolution optical tweezers. *DNA repair*, 20:58–70, 2014.

[82] Bettina Theissen, Anne R Karow, Jürgen Köhler, Airat Gubaev, and Dagmar Klostermeier. Cooperative binding of atp and rna induces a closed conformation in a dead box rna helicase. *Proceedings of the National Academy of Sciences*, 105(2):548–553, 2008.

[83] Denis Tsygankov, Martin Lindén, and Michael E Fisher. Back-stepping, hidden substeps, and conditional dwell times in molecular motors. *Physical Review E*, 75(2):021909, 2007.

[84] Yann R Chemla, Jeffrey R Moffitt, and Carlos Bustamante. Exact solutions for kinetic models of macromolecular dynamics. *The Journal of Physical Chemistry B*, 112(19):6025–6044, 2008.

[85] SC Kou, Binny J Cherayil, Wei Min, Brian P English, and X Sunney Xie. Single-molecule michaelis- menten equations, 2005.

[86] Martin Guthold, Xingshu Zhu, Claudio Rivetti, Guoliang Yang, Neil H Thomson, Sandor Kasas, Helen G Hansma, Bettye Smith, Paul K Hansma, and Carlos Bustamante. Direct observation of one-dimensional diffusion and transcription by escherichia coli rna polymerase. *Biophysical journal*, 77(4):2284–2294, 1999.

[87] Wei Cheng, Srikesh G Arunajadai, Jeffrey R Moffitt, Ignacio Tinoco, and Carlos Bustamante. Single–base pair unwinding and asynchronous rna release by the hepatitis c virus ns3 helicase. *Science*, 333(6050):1746–1749, 2011.

[88] Zhi Qi, Robert A Pugh, Maria Spies, and Yann R Chemla. Sequence-dependent base pair stepping dynamics in xpd helicase unwinding. *Elife*, 2:e00334, 2013.

[89] Ashley R Carter, Maasa H Seaberg, Hsiu-Fang Fan, Gang Sun, Christopher J Wilds, Hung-Wen Li, and Thomas T Perkins. Sequence-dependent nanometer-scale conformational dynamics of individual recbcd–dna complexes. *Nucleic acids research*, 44(12):5849–5860, 2016.

[90] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Mark Akeson, and Hongyun Wang. Dynamics of the translocation step measured in individual dna polymerase complexes. *Journal of the American Chemical Society*, 134(45):18816–18823, 2012.

[91] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Ashley Cox, Mark Akeson, and Hongyun Wang. Kinetic mechanism of translocation and dntp binding in individual dna polymerase complexes. *Journal of the American Chemical Society*, 135(24):9149–9155, 2013.

[92] Irina O Vvedenskaya, Hanif Vahedian-Movahed, Jeremy G Bird, Jared G Knoblauch, Seth R Goldman, Yu Zhang, Richard H Ebright, and Bryce E Nickels. Interactions between rna polymerase and the core recognition element counteract pausing. *Science*, 344(6189):1285–1289, 2014.

[93] F Ann Ran, Patrick D Hsu, Chie-Yu Lin, Jonathan S Gootenberg, Silvana Konermann, Alexandro E Trevino, David A Scott, Azusa Inoue, Shogo Matoba, Yi Zhang, et al. Double nicking by rna-guided crispr cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380–1389, 2013.

[94] Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

[95] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the minion nanopore sequencer. *Nature methods*, 12(4):351–356, 2015.

[96] G Adam and M Delbrück. Reduction of dimensionality in biological diffusion processes. *Structural chemistry and molecular biology*, 198, 1968.

[97] Cornelis Storm and PC Nelson. Theory of high-force dna stretching and overstretching. *Physical Review E*, 67(5):051906, 2003.

[98] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[99] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[100] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, pages 297–307, 1989.

[101] Manuel Barrio, André Leier, and Tatiana T Marquez-Lago. Reduction of chemical reaction networks through delay distributions. *The Journal of chemical physics*, 138(10):104114, 2013.

[102] Sam Walcott. The load dependence of rate constants. *The Journal of chemical physics*, 128(21):06B601, 2008.

[103] Mary L Boas. *Mathematical methods in the physical sciences*, volume 2. Wiley New York, 1966.

Appendix A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

A.1 Materials and Methods

Proteins: The M2-NNN-MspA protein6 was custom ordered from GenScript. Wild-type phi29 DNAP (833,000 U/ml; specific activity 83,000 U/mg) was obtained from Enzymatics or Epicenter. Hel308 was expressed using standard techniques by in-house facilities. Both phi29 DNAP and Hel308 were stored at -20 C until immediately before use.

DNA constructs: DNA oligonucleotides were synthesized at Stanford University Protein and Nucleic Acid Facility and purified at their facility using column purification methods. The oligo sequences are shown in table A.2. For both phi29 DNAP and Hel308 experiments, the nanopore read the same sequences for DNA threaded 5′ first. For phi29 DNAP experiments, sequences were previously used in [31]. In particular, DNA templates, primers and blocking oligomers were mixed at relative molar concentrations of 1:1:1.2 and annealed by incubating at 95 C for 3 min followed by slow-cooling to below 30 C.

To promote loading of Hel308 onto the template DNA strand, we annealed the template to a complement primer such that the template strand had an eight base 3′ overhang (A.7, A.2). A 5′ cholesterol on the complement strand promoted the DNA binding to the bilayer, and increased the interaction rate of DNA with the pore [96]. In solution, Hel308 binds to the 3′ overhang on the template strand and may begin to unwind the dsDNA in the 5′ direction. The 5′ end of the template DNA strand is drawn into the pore by the voltage, causing the complement strand to dissociate16. Hel308 bound to the DNA prevents complete translocation of the template strand through MspA17. Hel308 then functions as a translocase, drawing the ssDNA out of the nanopore in the 3′ direction back into the cis well. Figure A.7 illustrates the DNA translocase activity of DNA of through MspA by Hel308. We recorded

¿2000 current traces in various conditions, demonstrating enzymatic movement along DNA (A.2). With 180 mV applied, the currents were higher with Hel308 than with Phi29 DNAP because the buffer contained higher [KCl].

Operating Buffers: For phi29 DNAP experiments we used buffers of 300 mM KCl or asymmetric 150 mM cis KCl and 500 mM trans KCl, both with 10 mM HEPES at pH 8.0, 1 mM EDTA, 1 mM DTT and 10mM MgCl2. For Hel308 we used buffer at 400 mM KCl with 10 mM HEPES at pH 8.0, 1 mM EDTA, 1 mM DTT, and 10 mM MgCl2. Buffer [KCl] was higher than for phi29 DNAP experiments because the helicase operated better in higher salinity conditions.

Nanopore experiments: The experiments containing single M2-NNN MspA nanopores were established using thoroughly established techniques6,18. In short, we formed a lipid bilayer with 1,2-diphytanoyl-sn-glycerol-3- phosphocholine (Avanti Polar Lipids) across a horizontal 20 μM diameter aperture separating two 60 μL chambers containing our operating buffers. An Axopatch 200B or Axopatch 1B integrating patch clamp amplifier (Axon Instruments) applied a 180 mV voltage (unless otherwise noted) across the bilayer (trans side positive) and measured the ionic current through the pore. M2-NNN MspA was added to the grounded cis compartment to a final concentration of 2.5 ng/ml. Once a single pore inserted, as seen by a characteristic increase in the conductance, the buffer was replaced with MspA-free buffer to prevent additional pore formation. The DNA was added to the cis compartment to a final concentration of 10 nM. In a standard Phi29 DNAP experiment dCTP, dATP, dTTP and dGTP was added at the final concentrations of 100μM and Phi29 DNAP was added to a final concentration 20nM. In standard Hel308 experiments, our buffers of 400 mM KCl were premade with varying concentrations of ATP (10 μM, 20 μM, 50 μM, 250 μM, 500 μM, 1 mM, 3 mM). 1 mL of the chosen premixed solution was perfused into the cis chamber, ensuring the uniform concentration of ATP. In the Hel308 experiments, DNA was added to a final concentration of 10 nM and Hel308 to a final concentration of 100 nM. Unless otherwise mentioned, experiments were done at room temperature (23 $\pm$ 1 C).

Measurement of DNA position for Hel308 experiments: After scaling Hel308

levels and positioning them relative to the levels previously measured with phi29 DNAP, we found that all Hel308 levels lie along a spline interpolant between the phi29 DNAP current levels (Fig. 2.3c). Odd numbered Hel308 levels are close to previously observed phi29 DNAP current levels. Even numbered Hel308 levels lie along the interpolant somewhere in-between previously measured levels. As above in Figure 2.3d, we found the position of both the even and odd numbered Hel308 levels relative to the levels taken with phi29 DNAP.

Data acquisition and analysis: Data was acquired at 50 kHz with acquisition software written in LabView (National Instruments). Current traces were analyzed using custom programs written in Matlab (The MathWorks), Java and C. Collected data were box-filtered with at 10 point window and down-sampled to 5.0 kHz. DNA interactions and enzyme motor events were detected using previously described algorithms [27, 33, 32, 31, 26]. Ion current levels were selected automatically using the level finding algorithm used in [32]. using elements of the level finder more thoroughly described in [34]. Event counts and statistics are summarized in table A.2.

A.2 Experiment Statistics and DNA strands

Experiment Description	Enzyme motor	DNA sequence	[ATP] (μM)	Voltage (mV)	Temperature (°C)	Number of Events
ATP Titration	Hel308	A	10	180	22	44
ATP Titration	Hel308	A	20	180	22	23
ATP Titration	Hel308	A	50	180	22	102
ATP Titration	Hel308	A	250	180	22	63
ATP Titration	Hel308	A	500	180	22	212
ATP Titration	Hel308	A	1000	180	22	132
ATP Titration	Hel308	A	3000	180	22	34
Repetitive pattern	Hel308	C	500	180	22	30
ATP Titration	Hel308	B	50	180	37	58
ATP Titration	Hel308	B	100	180	37	27
ATP Titration	Hel308	B	500	180	37	38
ATP Titration	Hel308	B	1000	180	37	68
Voltage Repositioning	phi29 DNAP	B	--	140	22	70
Voltage Repositioning	phi29 DNAP	B	--	180	22	47
All sequence A phi29 data	phi29 DNAP	A	--	180	22	47*
All sequence B phi29 data	phi29 DNAP	B	--	various	22	233*
All sequence C phi29 data	phi29 DNAP	C	--	180	22	399*
All sequence A Hel308 data	Hel308	A	various	180	22	2110
All sequence B Hel308 data	Hel308	B	various	180	37	194
All sequence C Hel308 data	Hel308	C	various	180	22	61

Table A.1: Experiment Statistics For chapter 2

A summary of the statistics for all of the experiments performed in this study. Rows starting with ‘ATP titration‘ refer to experiments where we varied only [ATP] to determine the ATP dependent and independent steps, and the reaction kinetics of Hel308 during translocation of DNA. Rows starting with ‘repetitive pattern‘ refer to a sequence containing repeated current levels with high current difference. Rows starting with ‘Voltage repositioning‘ refer to phi29 DNAP experiments where the voltage was changed to study the effects of force repositioning the DNA within MspA. The final six rows summarize the total number of events for each of the DNA constructs examined using Hel308 or phi29 DNAP. Asterisks indicate that some data was used in [31, 33].

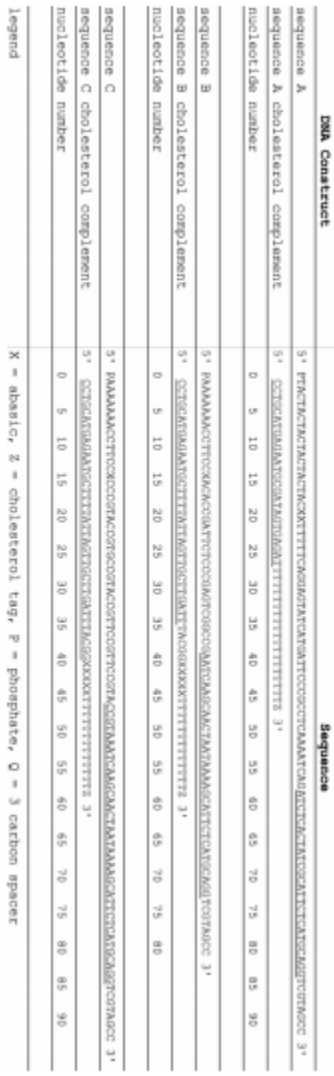

Table A.2: List of DNA Sequences Used in chapter 2

A list of the DNA strands and complements used in this study. Complementary sections are underlined. Sequence A was the sequence used in the main text.

A.3 Elongation of DNA

To test the hypothesis that ssDNA elongates within MspA with increasing force, we compared our results to the extensible freely jointed chain (Ex-FJC). The Ex-FJC is an experimentally validated model of DNA elongation under an applied force denoted F [97]. At forces in the 5-40 pN regime, the Ex-FJC gives the end-to-end distance of DNA, x, by the following expression:

$$x = L \cdot (1 - \frac{k_B T}{Fb}) \cdot (1 + \frac{F}{S}) \tag{A.1}$$

where L is the contour length of DNA, k_B is the Boltzmann constant, T is the temperature, S is the stretch modulus of ssDNA (800 pN) and b is the Kuhn length (1.45 nm)[54]. In our system, the end-to-end distance of DNA between the enzyme Phi29 DNAP and MspAs constriction, x, is fixed. The contour length, L, changes with different applied forces. We assume that the force on the DNA is proportional to the applied voltage $F = \alpha \cdot V$, giving:

$$x = L \cdot (1 - \frac{k_B T}{\alpha V b}) \cdot (1 + \frac{\alpha V}{S}) \tag{A.2}$$

At a different voltage $\beta \cdot V$, DNA is elongated by a different amount $\omega \cdot L$. ω is the ratio of contour lengths of the DNA between the enzyme and phi29 DNAPs constriction at the two different voltages. We substitute $V \to \beta \cdot V$ and $L \to \omega \cdot L$, in equation A.2 giving:

$$x = L \cdot \omega \cdot (1 - \frac{k_B T}{\beta \alpha V b}) \cdot (1 + \frac{\beta \alpha V}{S}) \tag{A.3}$$

Solving eqs. A.2 and A.3 for ω yields:

$$\omega = \beta(\frac{b\alpha V - k_B T}{b\alpha\beta V - k_B T}) \cdot (\frac{S + \alpha V}{S + \beta\alpha V}) \tag{A.4}$$

The fractional elongation ω can be recast as $\omega = \frac{N+\delta}{N}$, where δ is measured as in figure2.2d and N is the number of nucleotides between phi29 DNAP and MspAs constriction DNA at the initial force F. From [29] we estimate N = 12. We fit our data to eq. A.4, as shown in

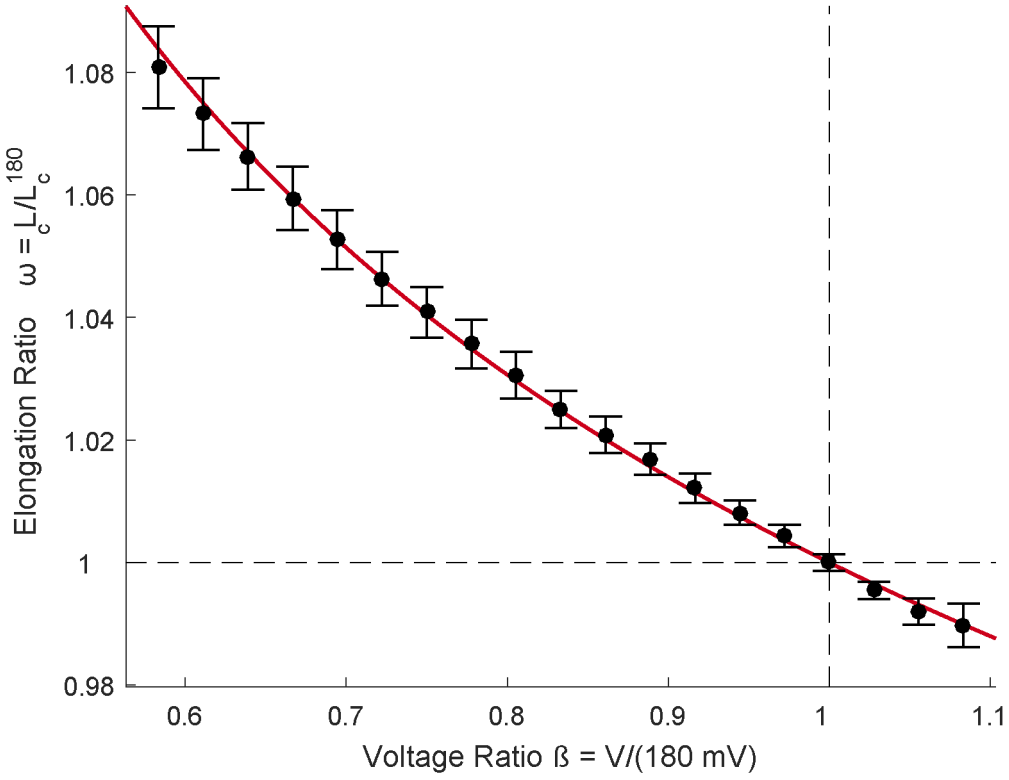

Figure A.1: DNA stretching in response to applied force

The elongation ratio ω as a function of the applied voltage. Error bars are S.E.

Figure

reffig:stretch, showing that a single parameter fit with $\alpha = 1.32 \pm 0.1 \frac{e^-}{nt}$ describes the data well. At 180 mV, using Eq. A.4, we estimate the DNA to be to be 92% of fully stretched.

Using Eq. A.4 with our data we find $F = \alpha \cdot V = 38 \pm 7pN$ at 180 mV. The uncertainty originates from (a) the measurements at different voltages and (b) allowing N to vary by +2 or -2. This force is higher than the anticipated forces of around 10-20 pN [61], requiring further exploration to directly calibrate the force applied to DNA within MspA using SPRNT. Assuming that SPRNT operates well under voltages from 80 mV to 240 mV, then the force range of SPRNT is roughly 15-50 pN.

A.4 Automatic Consensus Generation from a Reference

Off-pathway enzyme motor behaviors such as skips, backwards steps, and pausing complicated building an unbiased consensus of ion current levels. We developed a method that uses observed ion current levels and compares them to an ion current prediction to generate an improved consensus current level sequence. Our method uses the sequence alignment methods developed and described in [32]. In this alignment method, event level traces are aligned to a consensus level sequence, allowing for uncertainty in the consensus sequences.

We start with a set of current sequences from enzyme motor events, from levels found from raw data, and a prediction of current levels found from phi29 DNAP, where we expect a progression consistent with one level per moved nucleotide. We align each event to the prediction and record the resulting quality score of the alignment. A first generation consensus is constructed by replacing each level in the prediction with the median of the measured levels aligned to that position.

The first generation consensus then becomes the new prediction, and the process is repeated iteratively to produce an $(i+1)^{th}$ generation consensus. Figure A.2 shows the results of the algorithm on sequence A Note that in several places the consensus converged to a value different than the prediction.

The alignment quality is sensitive to inaccuracies of the predicted levels, and hence, we allow for uncertainty in these levels. In-line with simulated annealing techniques, the algorithm adds an uncertainty 'Temperature' when updating each alignment. This temperature is then 'cooled' after each iteration to help ensure that the consensus converges. We call the magnitude of the initial temperature, T_0. The temperature used for the i^{th} iteration is given by:

$$T(i) = T_0 - c \cdot (i-1) \tag{A.5}$$

where c is the cooling rate. The temperature is not allowed to go below 0. We used $T_0 = 10pA$, with $c = 1.5$ pA/iteration. We take into account the statistical spread of the individual

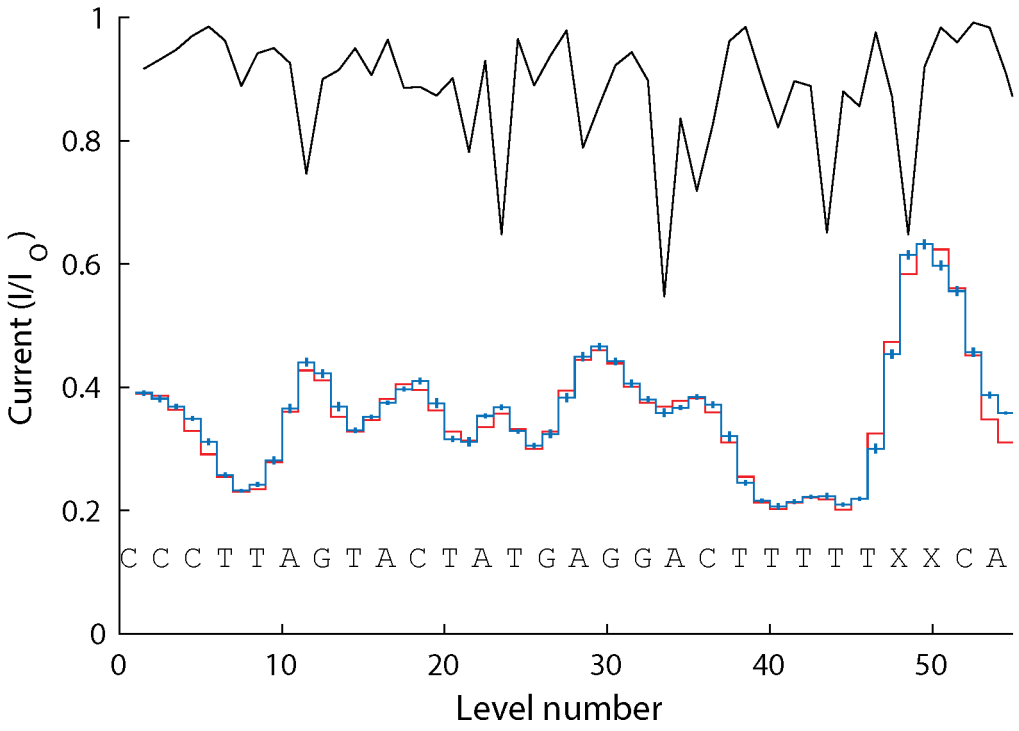

Figure A.2: Result of Consensus Algorithm

Predicted ion current (red) in normalized units, and results of the consensus algorithm (light blue) for DNA sequence A. The black line is the probability of observing an ion current level given that the DNA passed through the nanopore.

levels by adding the temperature in quadrature with the standard deviation of the measured values aligned to the L^{th} predicted level. The error on the L^{th} level of the i^{th} iteration of the consensus is given by:

$$s_L(i) = (T(i)^2 + \sigma_L^2)^{0.5}, \tag{A.6}$$

where σ_L is the standard deviation of all measured levels which aligned to the L^{th} predicted level.

Our prediction for Hel308 needed to reflect that there were two observed current levels for every nucleotide moved by the helicase. We had previously measured all possible four-nucleotide currents using phi29 DNAP [32]. We used the known sequences of our DNA strands to generate the prediction of the progression of current levels assuming a full nucleotide step as seen with phi29 DNAP. We fitted these predictions to a spline that was used to estimate the initial values at half nucleotide intervals between levels to generate a guess for the additional Hel308 levels.

The benefit of this algorithm is that it keeps the current levels in the preserved order and close to their correct sequence position. We considered several possible issues with the algorithm. Firstly, degenerate current regions (current levels separated by indistinguishable current differences) in the measured levels can lead to misalignments, and the shifting of levels from their correct positions. Second, there was a possibility that our consensus was biased towards the initial predicted current sequence. The initial prediction is important as we use the observed sequence predictions to estimate the distances the DNA moves under the control of the helicase. We ensured insensitivity of the consensus-building algorithm to the input initial predictions by taking initial predictions estimated with different step sizes from between 0.2 nt to 0.8 nt in steps of 0.1 nt. We found that the consensus converged to the same levels regardless of the initial step choice, suggesting minimal bias to our initial guess of 0.5 nt step sizes.

A.5 Hel308 Step Size Measurements

We find the position of a given sequence of DNA when translocated by Hel308 in comparison to the same sequence of DNA translocated by phi29 DNAP. First, we find a consensus of Hel308 current levels and a consensus of phi29 DNAP current levels for the same DNA strand (Appendix A.4). Next, we aligned the Hel308 current levels to the DNA sequence. For levels observed with Hel308 we generate a spline interpolant for even-numbered levels and a separate spline for odd-numbered levels. Next, a linear scale and offset was applied to the phi29 DNAP current level values to compensate for different salt conditions between the phi29 DNAP and hel308 experiments. For both the spline of even-numbered and of odd-numbered Hel308 levels, we shift the horizontal position of the spline curves, and take a sum of square differences between the splines of the Hel308 data and the spline for levels observed with phi29 DNAP. The shift leading to the smallest sum of square differences is taken as the DNA position for the set of Hel308 levels relative to the phi29 DNAP levels. The error on the DNA position measurements was calculated using the following Monte-Carlo simulation. 1000 perturbed sequences of the Hel308 and phi29 DNAP levels were produced randomly using the known errors on the current levels. The above calculation was repeated for each of the 1000 sequences to generate a distribution of DNA position measurements. The standard deviation of this distribution is the error on the DNA position measurements. For each set of levels taken with Hel308 (even or odd numbered levels) we found the mean step position relative to phi29 DNAP. For DNA sequences A, B, C we found the distance between steps to be $0.54 \pm 0.04\ nt$, $0.54 \pm 0.04\ nt$ and $0.44 \pm 0.05\ nt$, respectively.

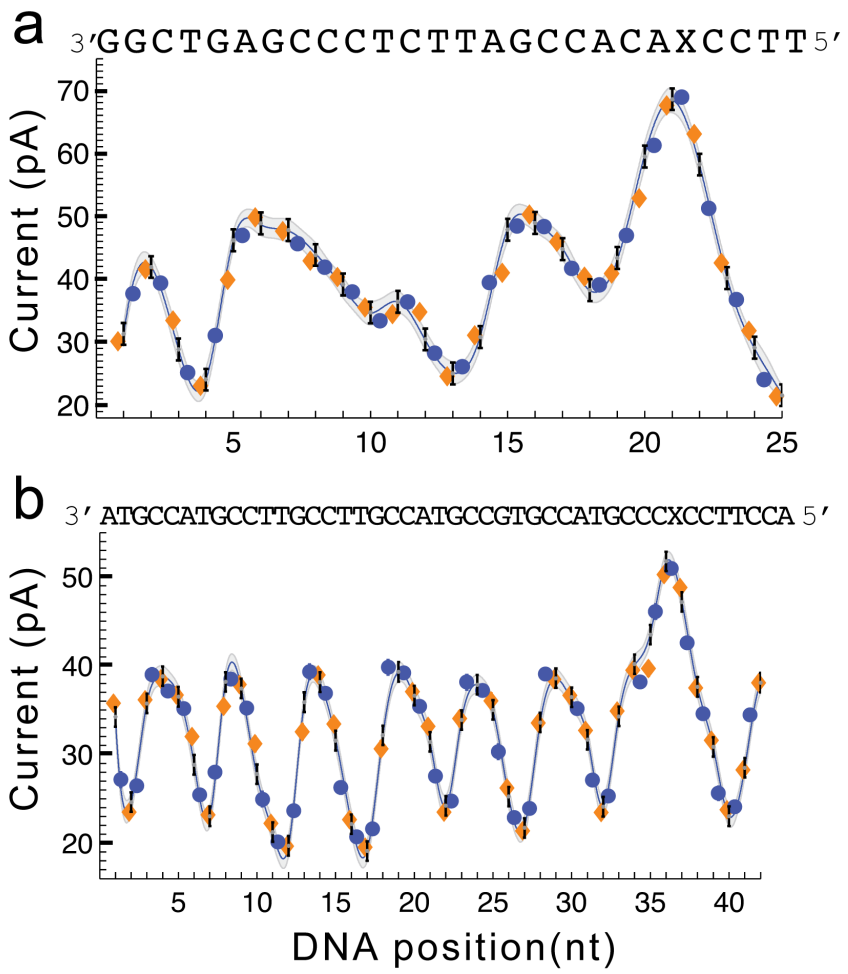

Figure A.3: Measuring Hel308 Step Size

Figure caption is the same as Fig. 2.3c,d but for DNA sequences B and C displayed in (a) and (b), respectively. For both (a) and (b) the gray curve represents the spline of the levels observed with phi29 DNAP (black points) moving the DNA through MspA. Means of current levels recorded with Hel308 actuated DNA movement (orange and blue symbols) were scaled to match the spline of phi29 DNAP levels. Points indicated with orange diamonds or blue circles were horizontally offset in order to best match the spline of levels taken with phi29 DNAP. Levels that were found to be depend on [ATP] are shown with gold diamonds, and levels that are independent of [ATP] are shown with blue circles.

A.6 ATP Titration of Hel308

To examine the nature of the two states of the Hel308 hydrolysis cycle we systematically varied the ATP concentration from 10 μM to 3 mM. We found, consistent with figure 2.3f, that the even-numbered levels did not change in duration with ATP concentration and that the odd levels did. Supplemental figure A.6 shows the median reaction time, averaged over all levels as a function of the 1/[ATP] for the odd (gold) and even (blue) steps. We fit these curves to lines and find that $\tau_{1/2} = [(4.1 \pm 0.4)\mu M/[ATP] + (0.05 \pm 0.01)]s$ for the [ATP]-dependent step, and $\tau_{1/2} = [(0.08 \pm 0.8)\mu M/[ATP] + 0.16 \pm 0.01]s$ for the [ATP]-independent step. Because the error is much larger than the value of the slope, we conclude that this step does not depend on [ATP].

An inset shows the inverse plot of rate vs. [ATP]. We find that the ATP dependent step follows the Michaelis-Menten equation ($rate = \frac{V \cdot [ATP]}{K + [ATP]}$), with $V = 15.5 \pm 3.5 s^{-1}$ and $K = 92 \pm 22 \mu M$.

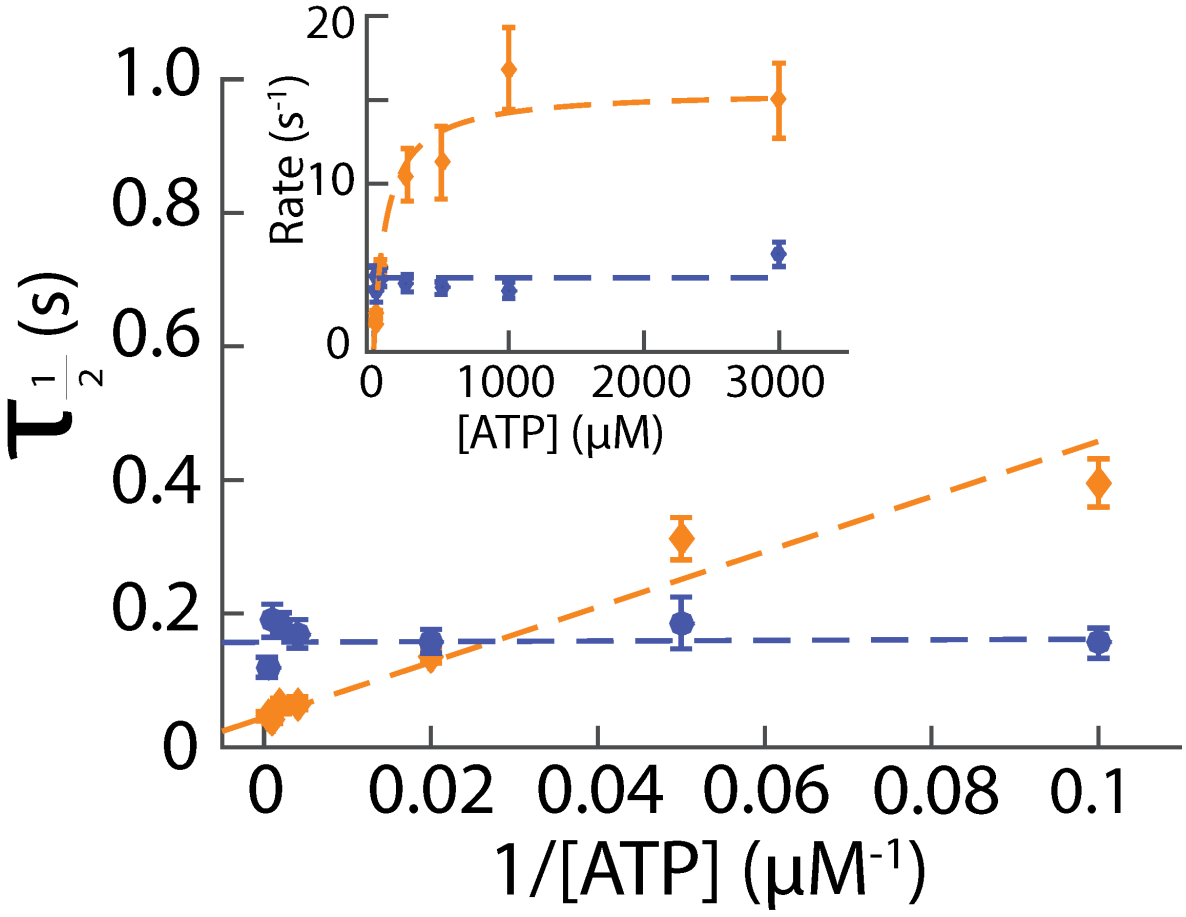

Figure A.4: ATP Titration of Hel308

The median duration for the [ATP]-dependent step, averaged over all [ATP]-dependent levels (gold) and the [ATP]-independent step, averaged over all [ATP]-independent levels (blue) as a function of 1/[ATP], with best fit lines drawn as dashed lines over both. The inset shows the inverse plot of rate (rate $=\frac{ln(2)}{\tau_{1/2}}$) vs. [ATP].

A.7 Proposed 2-Step Mechanism and Hel308 Translocase Experiment

In this section we suggest a mechanism for why we observe two steps with Hel308 and MspA.

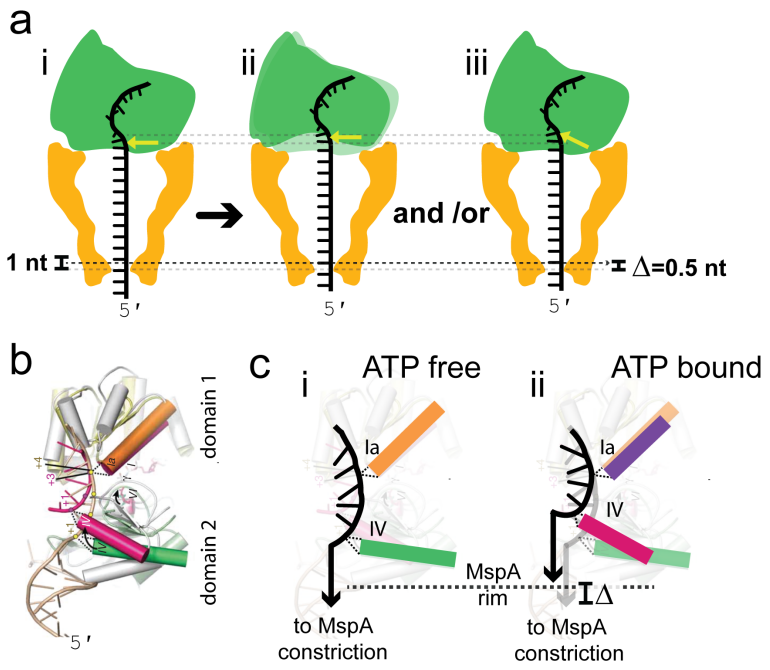

Figure A.5: Proposed Mechanism of 2-step Hel308 Translocase

(a) Illustration of DNA (black) moving within MspA (gold) during Hel308 (green) translocase activity. Hel308 starts in the conformation shown in (i). When ATP binds, the physical structure of Hel308 changes, altering how it sits on the rim of MspA, as in (ii), and/or repositioning the DNA within the Hel308, as in (iii). Yellow arrows indicate a DNA binding motif within the enzyme that can move the DNA relative to MspAs constriction. (b) The image from Buttner et al compares domains 1 and 2 between two crystal structures of Hel308 and another Ski2 like helicase, one without ATP bound, another with ATP bound. Buttner et als analysis indicates that ATP binding induces a conformational rotation of domain 2 by 20 degrees, and consequently moves the DNA binding motif IV closer to domain 1. (c) We take the illustration in (b) and highlight the DNA (black) and motifs Ia helix (orange) and motif IV helix (green), for the ATP unbound helicase (i) and for the ATP bound helicase (ii). The remaining 5′ end of the DNA is threaded through MspAs constriction. Upon ATP binding, motif IV helix (magenta) repositions the DNA upwards towards domain 1, while the DNA-binding domain Ia helix (purple) remains nearly unmoved. Because the anticipated contact points between the helicases and MspAs rim (horizontal dashed line) does not change the position of the helicase considerably, the helicase will move DNA relative to MspA constriction by an amount Δ. See Appendix A.5 for more information on step size measurement. The remainder of the ATP hydrolysis cycle returns the Hel308 to its original conformation while completing the translocation of the DNA by one nucleotide through the pore.

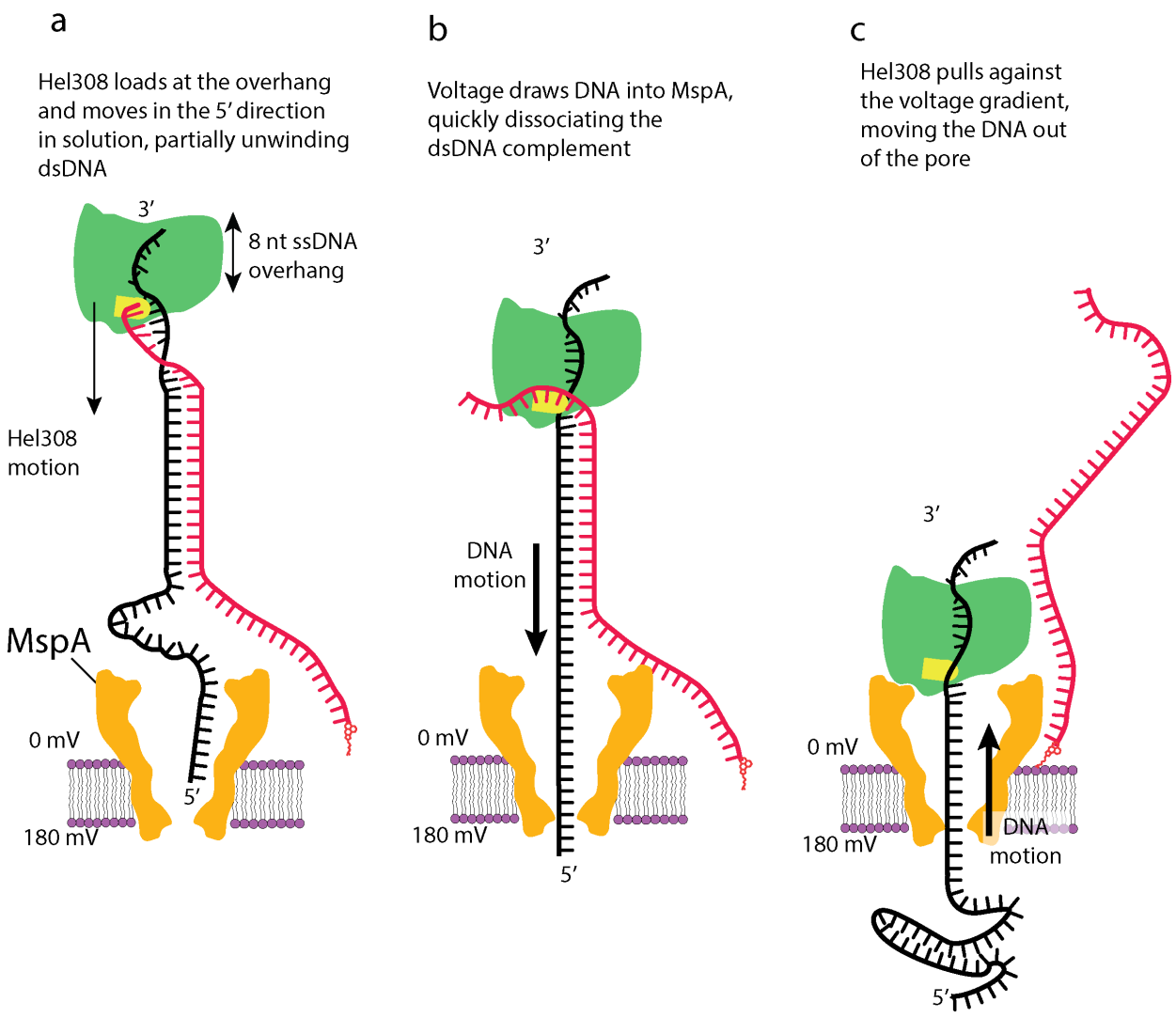

Figure A.6: Hel308 Translocase Experiment

An illustration of DNA being moved through MspA (gold) by the helicase Hel308 (green) in a lipid bilayer (purple). (a) Hel308 loads onto the exposed 3′ end of the template DNA strand (black). The complement DNA strand (red) was designed to leave an 8 base overhang on the template strand 3′ end on which the helicase loads. A cholesterol is attached to the 3′ end of the compliment DNA strand to concentrate DNA onto the bilayer. (b) Hel308 partially unwinds the complement DNA strand until the voltage quickly dissociates the remainder of the complement DNA strand from the 5′ end while pulling the template strand through MspA. (c) Hel308 functions as a 3′ to 5′ ssDNA translocase, drawing the DNA out of the pore. As Hel308 progresses along the DNA we observe discrete changes in current, indicative of Hel308 translocase activity.

Appendix B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

B.1 Sources of Ion Current Modulation in Nanopore Experiments

Understanding the ion current through the nanopore is essential in properly analyzing enzyme kinetics with SPRNT. There are several potential sources of ion current modulation in nanopore systems:

1. Enzyme / DNA motion
2. Access resistance changes caused by the enzyme resting close to MspA
3. Fluctuations caused by contamination of ions other than K^+ and Cl^-
4. Interactions between the DNA bases and MspA

Our goal is to decouple current changes caused by enzyme activity from the other sources of current modulation.

In the Hel308 ion current traces, we observed that many ion current steps (both [ATP]-dependent and [ATP]-independent) tended to have short-lived decreases in ion current amplitude (5-50 ms) that could not be associated with any Hel308 ion-current amplitude, before returning to the previous ion current step (Fig. B.1). We call such ion-current amplitudes 'flickers'. Because flickers occurred as only downwards spikes in the current trace, it is also unlikely that flickers are caused by DNA motion through the pore, as we would expect this to lead to both current increases and decreases [51]. To test if flickers were caused by ion contamination or interactions between the DNA bases and MspA, we performed an experiment in which we placed both phi29 DNA polymerase[31] and Hel308 into the reaction volume. Flickers were observed only in the translocation events with Hel308, suggesting that Hel308 is required for flickering, and thus flickers are not an effect of the ion contamination

or MspA-DNA interactions. We thus attribute flickering behavior to a transient access resistance caused by interactions between Hel308 and MspA. We conclude that the flickers are not caused by enzymatic activity. Therefore, when examining dwell times we must include flickers together with the associated step.

We analyzed the data by performing sequence alignment of the current amplitudes for each event to a reference consensus [51, 32]. Flickers tended to be aligned to an incorrect step, so we corrected each alignment manually using a custom-made GUI.

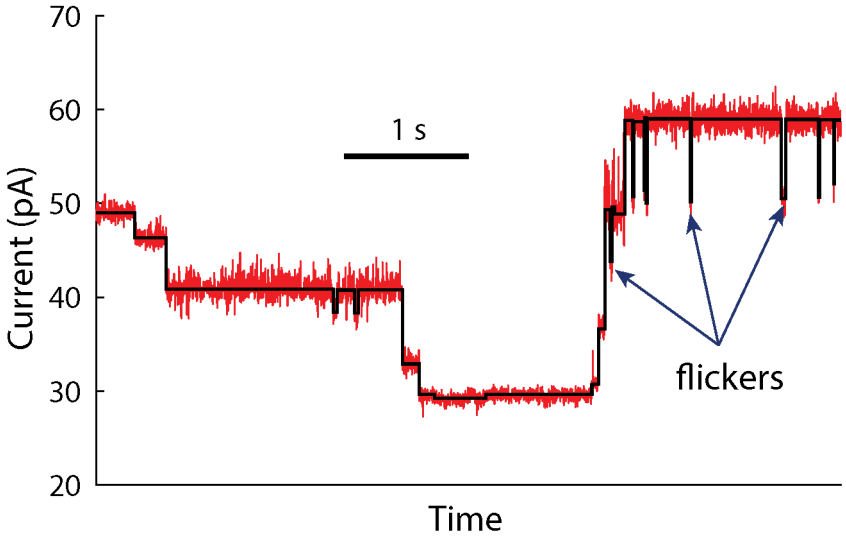

Figure B.1: Analysis of Ion Currents

An ion current trace (red) with automatically found current steps (black) plotted on top. The arrows indicate several examples of flickers, defined as short, transient decreases in the ion current that cannot be associated with any Hel308 ion current steps.

B.2 Conversion of Ion Current to DNA Position

As an enzyme walks along a DNA strand and feeds it into the MspA pore, we observe a series of discrete ion current steps (Fig. B.2a). Previously, we found that these ion current steps observed during nanopore sequencing are a discrete sampling of a smooth underlying curve of ion current versus DNA position [51]. This smooth curve is what one would observe if they were to smoothly feed DNA through the pore and plot the measured ion current vs. the DNA position within the pore. Using alignment algorithms similar to Needleman-Wunch alignment [32, 98, 99] this smooth curve can also be used as a direct mapping from measured ion current to DNA position within the pore.

To make this conversion, one needs to know the underlying smooth current vs. position curve for the particular sequence of DNA used. Previously, we found that the Hel308 helicase steps in two approximately half-nucleotide steps on ssDNA[51] ([ATP]-dependent step forwards to ATP-independent: 0.55 ± 0.03 *nt* and [ATP]-independent forwards to [ATP]-

dependent: $0.45 \pm 0.03\ nt$). In this manuscript, we approximate this smooth underlying curve with a spline curve determined from the consensus of current values observed while Hel308 helicase steps DNA through the pore. Consensus DNA positions are spaced according to our previous measurement (0, 0.45 nt, 1.0 nt, 1.45 nt, 2.0 nt, ...; fig B.2b). A given raw SPRNT current measurement can then be converted from ion current to position via the following steps:

1. Find and extract average ion current values for steps in the data (Fig. B.2a). This is described in detail in [50].

2. Align extracted ion current step means to the previously measured consensus. For alignment we use a dynamic programming algorithm similar to Needleman-Wunch alignment [32, 98, 99]. For a detailed explanation of alignment of nanopore currents, see Laszlo 2016, Appendix C [50]. Average ion current steps for 20 example events have been aligned to the consensus in Figure B.2b.

3. Use the alignment from (2) to match each ion current datapoint from the ion current measurement to the corresponding DNA position. Bulk alignment of ion current steps to the consensus provides initial context that allows individual ion current datapoints to be matched to the underlying smooth curve. Because the measured currents are not unique to a particular position (i.e. multiple positions along the DNA result in similar or identical ion current measurements), bulk alignment of current steps allows us to determine where on the spline to look for a corresponding ion current/position pair. Matching is done via a T-test comparison of each measured current value to all spline current values that lie within 3 nt of the bulk-aligned position. The spline has some uncertainty in it because there is variability in the observed currents from measurement to measurement (see aligned levels in Fig B.2b), this uncertainty is used as the standard deviation, σ, in the T-test. For each measured ion current datapoint, the T-test yields a liklihood of match for each current/position pair of the spline (within 3 nt of the

bulk alignment). It sometimes occurs that there are two or three best possible current matches. This happens when the measured enzyme step is close to a peak or trough in the underlying smooth curve, thus a measured current point can match spline current values to the left and right of the peak/trough. We resolve this ambiguity by using our knowledge of the overall alignment to assign a prior probability to the T-test output. We thus multiply the position likelihood scores by an assigned prior probability that is a normalized Gaussian of width $\sigma = 0.7\ nt$, centered at the position of the alignment. Figure B.2 panels c, d, and e show in schematic form how ion current is transduced into position using the smooth current vs. position curve.

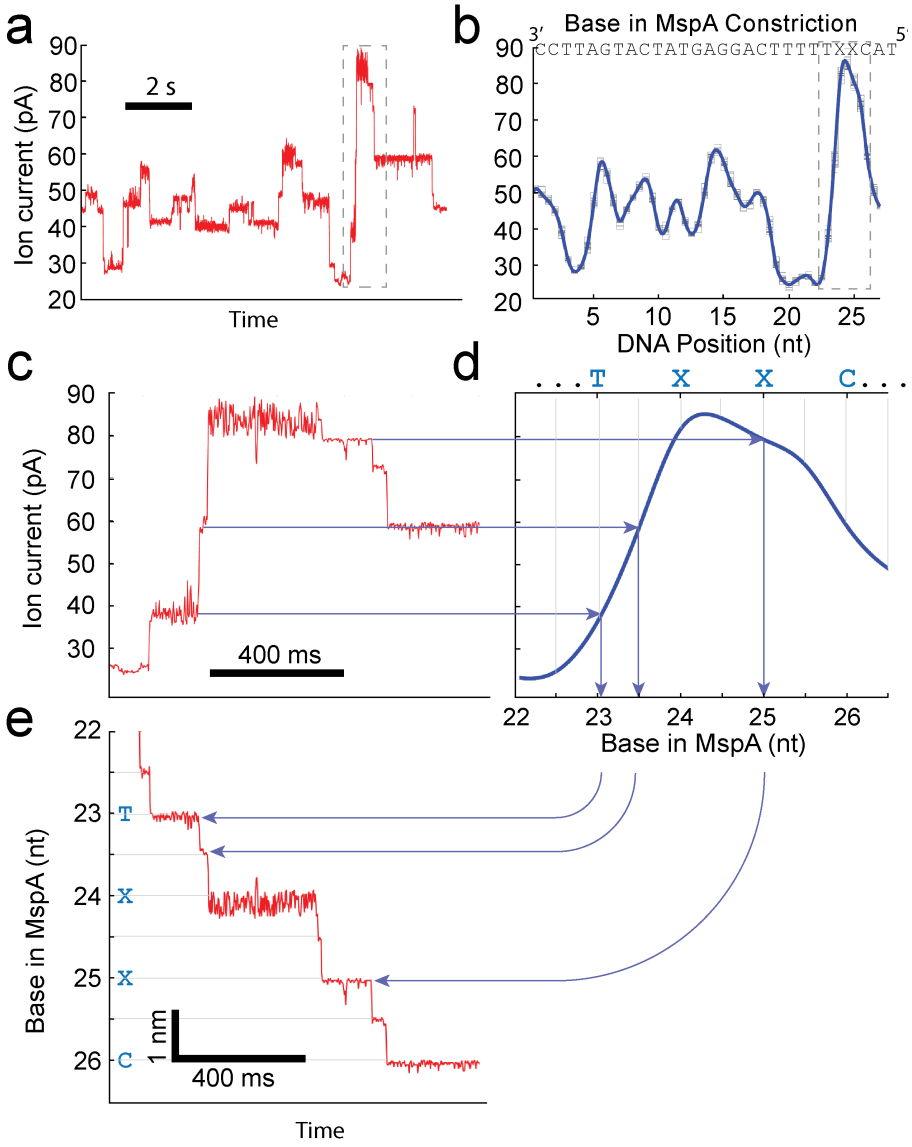

Figure B.2: Conversion of ion current to DNA position

(a) Raw data trace of ion current versus time for the same event shown in figure 1c. (b) Smooth ion current versus DNA position curve, constructed by averaging together many reads like those in (a)[50]. The DNA sequence in the MspA constriction is displayed above, with 'X' indicating an abasic site. (c-e) Each data point in the ion current versus time trace (c) is mapped to the underlying smooth curve (d) as indicated by the arrows, thereby determining the DNA position versus time (e).

Experiment	$[ATP](\mu M)$	$[ADP](\mu M)$	$Voltage(mV)$	$Temperature(°C)$	N_{events}
ATPt	10	0	180	22	34*
ATPt,ADPi,Voltage II	50	0	180	22	47*
ATPt	100	0	180	22	38
ATPt	250	0	180	22	59*
ATPt,Temperature,Voltage I	500*	0	180	22	36*
ATPt	1000	0	180	22	99*
ATPt	2000	0	180	22	35
ATPt	3000	0	180	22	57*
ADPi	50	10	180	22	22
ADPi	50	25	180	22	44
ADPi	50	50	180	22	30
ADPi	50	100	180	22	32
ADPi, Ratio [ATP]:[ADP]=1:4	50	200	180	22	27
Ratio [ATP]:[ADP]=1:4	300	1200	180	22	15
Ratio [ATP]:[ADP]=1:4	500	2000	180	22	33
Ratio [ATP]:[ADP]=1:4	700	2800	180	22	34
Voltage I	500	0	140	22	37
Voltage I	500	0	160	22	27
Voltage I	500	0	200	22	41
Voltage I	500	0	220	22	37
Voltage I	500	0	240	22	55
Voltage I	500	0	260	22	37
Voltage I	500	0	280	22	35
Voltage II	50	0	140	22	34
Voltage II	50	0	220	22	45
Voltage II	50	0	260	22	26
Temperature	500	0	180	28	17
Temperature	500	0	180	34	65
Temperature	500	0	180	45	28
Ratio [ATP]:[ADP]=1:2	10	20	180	37	39
Ratio [ATP]:[ADP]=1:2	50	100	180	37	30
Ratio [ATP]:[ADP]=1:2	100	200	180	37	22
Ratio [ATP]:[ADP]=1:2	250	500	180	37	40
Ratio [ATP]:[ADP]=1:2	500	1000	180	37	24
Ratio [ATP]:[ADP]=1:2	750	1500	180	37	48
Ratio [ATP]:[ADP]=1:2	1500	3000	180	37	27

Table B.1: Experimental conditions and number of Hel308 events

The experiments tags are as follows. ATPt refers to experiments in which only the ATP concentration was varied. An asterisk indicates some data was used in a previous publication [51]. ADPi refers to experiments in which the ADP concentration was varied at $[ATP] = 50\ \mu M$. Voltage I and Voltage II refer to experiments in which the voltage was varied while maintaining $[ATP] = 500\ \mu M$ and $[ATP] = 50\ \mu M$, respectively. Ratio refers to experiments in which we maintained a constant ratio of $[ATP] : [ADP] = 1 : 4$ or $[ATP] : [ADP] = 1 : 2$. Temp refers to experiments in which the temperature was varied at $[ATP] = 500\ \mu M$. In total $N = 1357$ single-molecule data traces were obtained for this study.

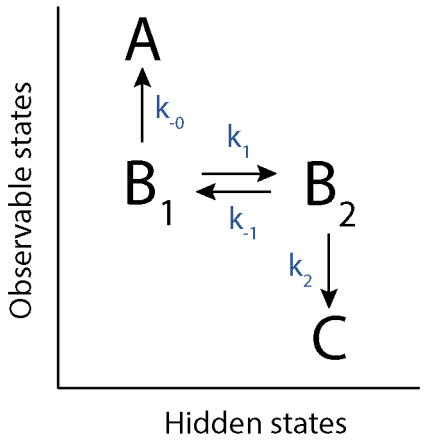

Figure B.3: A simple kinetic model to illustrate the master equation.

B.3 The Master Equation

To analyze different kinetic models at the single-molecule level we use the master equation formalism[43], which describes the way that probability flows between different enzyme states. The master equation can be used to answer questions such as: what is the average dwell time of an observable state as a function of the substrate concentration? What is the probability that the motor enzyme steps backwards? What is the distribution of dwell times for $f|f$ steps? Consider the toy model shown in figure B.3, a four state model consisting of chemical states A, B_1, B_2, and C which transition between one another. Transitions between A and B_1, and B_2 and C result in a change in ion current signal, while transitions between B_1 and B_2 are hidden. Assume that at time 0 (i.e. the start of a new DNA position measurement) the enzyme is in state B_1. The rate of change of the probability that the enzyme occupies state B_1 is obtained by summing the rate of probability flow into state B_1, and subtracting the rate of probability flow out of state B_1:

$$\frac{dp_{B_1}}{dt} = p_{B_2} \cdot k_{-1} - p_{B_1} \cdot (k_{-0} + k_1) \tag{B.1}$$

Equations of this form can be written for states, A, B_2 and C as well. Then we note that these

are a linear system of differential equations, which allows us to write the matrix equation:

$$\frac{d\vec{p}}{dt} = \frac{d}{dt}\begin{bmatrix} p_A \\ p_{B_1} \\ p_{B_2} \\ dp_C \end{bmatrix} = M \cdot \vec{p}(t) = \begin{bmatrix} 0 & k_{-0} & 0 & 0 \\ 0 & -k_{-0} - k_1 & k_{-1} & 0 \\ 0 & k_1 & -k_{-1} - k_2 & 0 \\ 0 & 0 & k_2 & 0 \end{bmatrix} \cdot \begin{bmatrix} p_A \\ p_{B_1} \\ p_{B_2} \\ p_C \end{bmatrix}. \tag{B.2}$$

equation B.2 is called the 'master equation' for this system. The entries of $\vec{p}(t)$ are the probabilities that the individual states are occupied at time t and M is the called 'connection matrix'. Diagonal entries of M are outflow rates from a given state, while off diagonal entries M_{ij} are the rate of probability flow from state j into state i. The columns of M must always sum to 0 to conserve probability. In this example we start in state B_1, therefore the initial conditions for equation B.2 are:

$$\vec{p}(t=0) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{B.3}$$

The master equation together with the initial conditions completely specify the system. The solution to equation B.2 is easily written in terms of the eigenvalues of the connection matrix:

$$\vec{p}(t) = \Sigma_i c_i \cdot \vec{\xi_i} \cdot exp(\lambda_i \cdot t), \tag{B.4}$$

where λ_i and $\vec{\xi_i}$ are the eigenvalues and eigenvectors of M, and c_i are the coefficients of integration. The c_i can be solved according to the initial conditions:

$$\vec{c} = V^{-1} \cdot \vec{p}(t=0) \tag{B.5}$$

where V is a matrix whose columns are the $\vec{\xi_i}$. In principle, we have fully solved the problem, however for complex kinetic models with many states and parameters the eigenvalues

cannot be analytically solved. Thus using equation B.4 can be difficult. In the remainder of this supplement we will use several different techniques to analyze solutions to the master equation, such as the steady-state approximation, numerical solutions, direct integration and laplace transform.

It is important to discuss here how this formalism connects to experimental data. For example, when displaying the kinetic model in figure B.3 why did we not include the transition $A \overset{k_0}{\rightarrow} B_1$ and $C \overset{k_{-2}}{\rightarrow} B_2$? These terms are certainly important, but not relevant to the experimental question we are asking: what is that probability distribution of dwell times for the enzyme to go between observable states given that we begin in state B_1? This is an example of the 'first-passage time' problem. If we were to include $A \overset{k_0}{\rightarrow} B_1$ in this diagram and the master equation, then we would be including information from multiple different observable states into our model, which complicates interpretation of the data. The term $A \overset{k_0}{\rightarrow} B_1$ would be included in discussing the master equation for a different observable state.

B.4 Statistical Analysis of Michaelis-Menten Parameters for f|f [ATP]-dependent Steps

We tested whether the variation among measured Michaelis-Menten parameters $V_{f|f}$ and $K_{f|f}$ (MT eq. 1, figure B.4) could have been produced by statistical fluctuations. We asked the following statistical question: what is the probability of observing the joint distribution of experimentally measured $V_{f|f}$ and $K_{f|f}$, given that $V_{f|f}$ and $K_{f|f}$ do not depend on DNA position (the null-hypothesis). We generated data for the null-hypothesis by placing all of the data for each [ATP]-dependent step (all half-integer DNA positions) at a given [ATP] into a single bin. The number of data points in a single concentration i is N^i_{total}. Let the average number of times a given DNA position is measured at a concentration i be N^i_{ave} (table B.2). For each concentration we drew N^i_{ave} measurements of the dwell time at random from the N^i_{total} measurements. We took the mean and standard deviation of the mean of each of these bins, and performed a weighted fit to the Michaelis-Menten equation, extracting the values of $V_{f|f}$ and $K_{f|f}$. These parameters represent 1 Monte Carlo sample in our null hypothesis. We repeated this process 10^5 times, which is justified because $10^5 \ll Choose(N^i_{total}, N^i_{ave})$ for each experiment, ensuring that each Monte Carlo sample is independent. Figure B.5 shows the joint distribution of $K_{f|f}$ and $V_{f|f}$ for both the Monte Carlo samples and the measured values (blue and red, with solid and dashed black lines representing the error ellipses, respectively). It is clear that many of the experimental data points could not have been produced randomly at confidence $p < 10^5$. As such, we reject the null hypothesis that the measured distribution of $K_{f|f}$ and $V_{f|f}$ could have been randomly produced at confidence $p \ll 10^5$.

$[ATP](\mu M)$	10	50	100	250	500	1000	2000	3000
N^i_{total}	424	1807	576	1159	493	1658	541	944
N^i_{ave}	18	75	24	48	21	69	23	39

Table B.2: Number of Measurements of f|f [ATP]-dependent steps

The total number of measured f|f steps for the [ATP]-dependent step (N^i_{total}) and average number of measurements for each step (N^i_{ave}) at the given ATP concentration.

B.5 Analysis of Probability of Backwards Steps

We define the probability of a backwards step for a DNA position j as:

$$p_{back,j} = \frac{N_{b|f,j}}{N_{b|f,j} + N_{f|f,j}}, \tag{B.6}$$

where $N_{f|f,j}$ and $N_{b|f,j}$ are the observed number of forwards and backwards steps following forwards steps for step j. For figures 4 and 6 in the main text we average over all DNA positions so that

$$p_{back,condition} = \frac{\Sigma_j N_{b|f,j}}{\Sigma_j N_{f|f,j} + N_{b|f,j}}. \tag{B.7}$$

The error on this measurement is calculated using the binomial distribution:

$$\delta p_{back} = \sqrt{\frac{p_{back} \cdot (1 - p_{back})}{N_{f|f} + N_{b|f}}}. \tag{B.8}$$

Figure 3.5 shows the probability of a backstep for the [ATP]-independent step at different DNA positions, averaged over each ATP and ADP titration experiment.

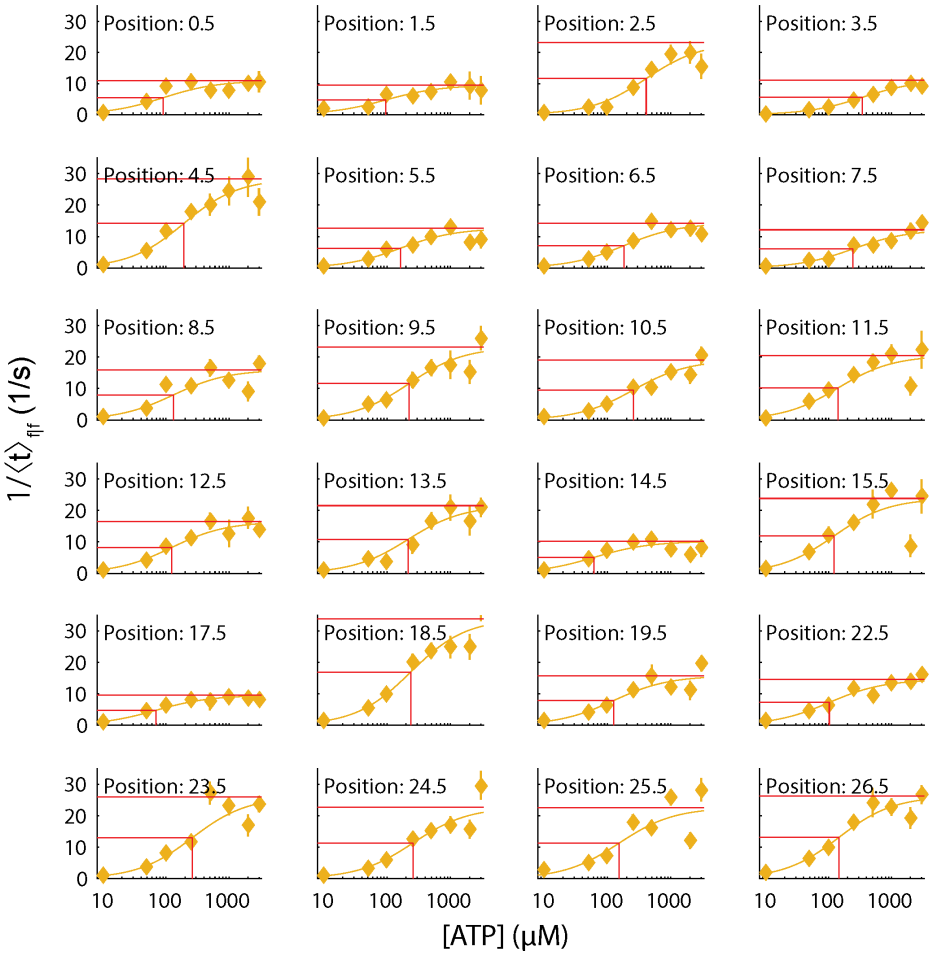

Figure B.4: ATP Dependence of All f|f [ATP]-dependent Steps

The rate of [ATP]-dependent f|f steps versus [ATP] at several DNA positions (positions 16.5 20.5 and 21.5 are ommitted due to degenerate ion current signals). The best fit of the Michaelis-Menten equation to the data is plotted on top (yellow line). The red line indicates the best fit value of the maximum rate of reaction ($V_{f|f}$). The horizontal blue line corresponds to $V_{f|f}/2$, with the vertical blue line showing the position of $K_{f|f}$.

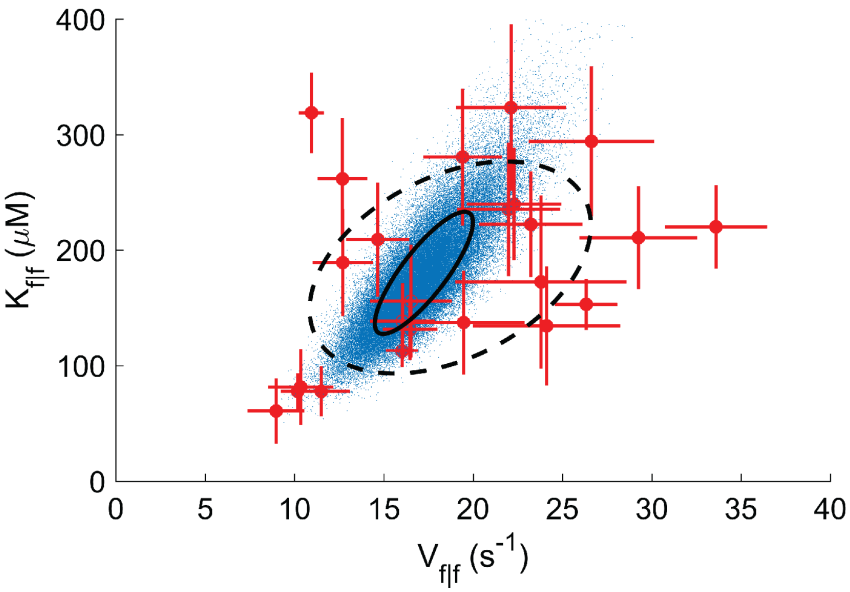

Figure B.5: Analysis of Michaelis Parameters

The distribution of Michaelis parameters for [ATP]-dependent steps (red). Crosses are the 1 S.E.M. measured error. Each blue dot is a Monte Carlo simulation, generated by taking the data from each step, drawing randomly and fitting a Michaelis Menten equation to the resulting mean values. The solid and dashed black curves are error ellipses for the Monte Carlo simulation and the measured data, respectively.

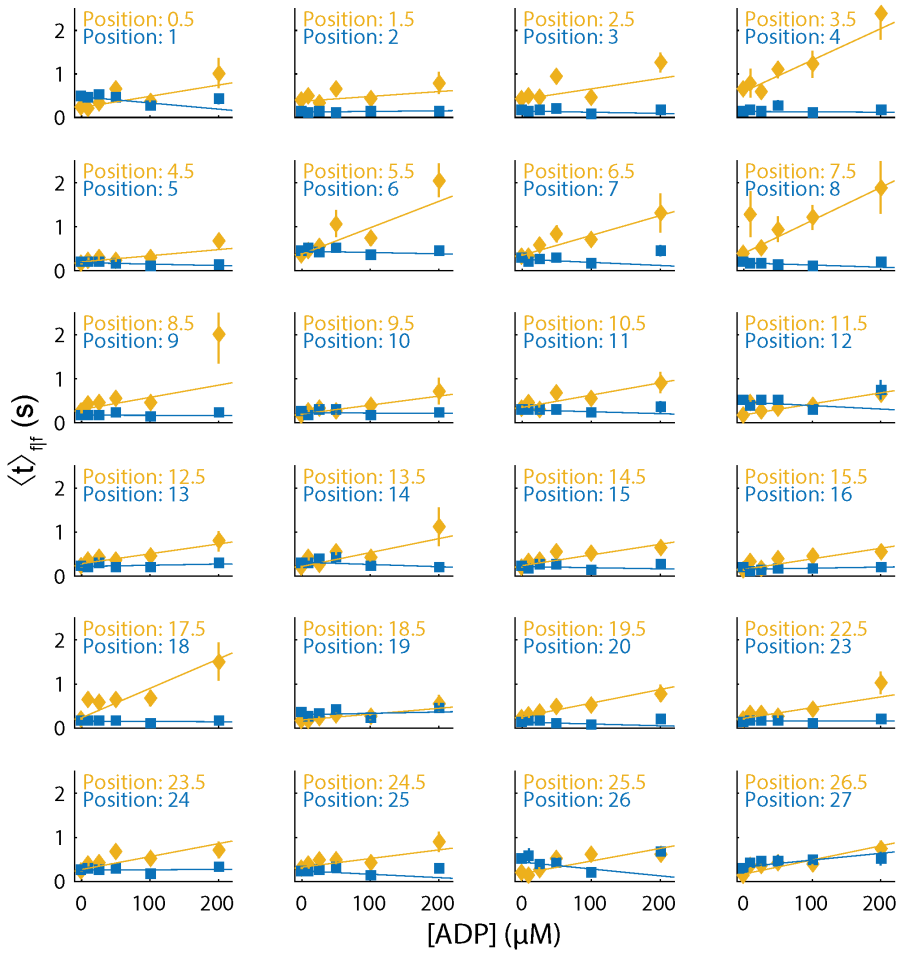

Figure B.6: ADP dependence of All f|f [ATP]-dependent Steps

The average dwell time of f|f [ATP]-dependent steps (yellow, half-integer DNA positions) and [ATP]-independent steps (blue, integer DNA positions) as a function of the [ADP]. The best linear fits are plotted on top. The average dwell time of the [ATP]-dependent step increases linearly with [ATP], while the average dwell time of the [ATP]-independent step does not depend on [ADP].

B.6 Comparing f|f and f|b [ATP]-dependent steps

To compare f|f to f|b [ATP]-dependent steps, we analyzed only those DNA positions in which the following [ATP]-independent step went backwards sufficiently often to accumulate enough statistics to analyze the ATP dependence of f|b [ATP]-dependent steps. Figure 3.6a shows the rate of reaction for f|f and f|b [ATP]-dependent steps. Each of these curves are nearly identical, and produce similar values of V and K when fit to the Michaelis-Menten equation. The expressions for V and K for $f|f$ and f|b steps from Model1 are (for derivations see Discussion B.7, B.10):

$$V_{f|f} = k_2 \cdot \frac{k_D}{k_2 + k_D}, V_{f|b} = k_2 \tag{B.9}$$

$$K_{f|f} = \frac{k_{-T} + k_2}{k_T} \cdot \frac{k_D}{k_2 + k_D}, K_{f|b} = \frac{k_{-T}}{k_T} \tag{B.10}$$

where the subscript indicates the step type. If we have that $k_D \gg k_2$ then $V_{f|f} = V_{f|b}$, and if we additionally have that $k_{-T} \gg k_2$ then $K_{f|f} = K_{f|b}$. Evaluating the model parameters for the [ATP]-dependent step confirms this to true in most cases (Table B.6). In some cases k_{-T} is similar in value to k_2, but the errors on the fit value of K do tend to be fairly large ($\approx 20 - 40\%$ for individual steps), so it may be difficult to distinguish differences between $K_{f|f}$ and $K_{f|b}$.

Figure 3.6b shows the dwell time distributions for the DNA positions (2.5, 7.5, 9.5 and 24.5) at subsaturating ($[ATP] = 50\ \mu M$, top row) and saturating ($[ATP] = 1\ mM,\ 2\ mM,\ 3\ mM$, bottom row) [ATP]. We used the 2-sample KS test to evaluate the similarity of each of the [ATP]-dependent f|f and f|b dwell time distributions (Fig. 3.6), and found that each pair of histograms was statistically indistinguishable ($p > 0.05$ for each pair of histograms). These results can be explained by one the following two arguments:

1. If the preceding b|f [ATP]-independent step is an off-pathway backwards step resulting from an unproductive forwards step, then we would expect that the initial conditions do not modify the kinetics, because the initial conditions are actually unmodified.

2. If the preceding b|f [ATP]-independent step is an on-pathway backwards step, then because the ATP and ADP off rates (k_{-T} and k_D) are large when compared with k_2, we would expect that f|f and f|b steps both effectively start in the free enzyme state (unbound Hel308), which would lead to similar kinetics.

It is impossible to distinguish between which of these cases is actually occurring due to the similarity of b|f and f|f [ATP]-independent steps (Fig 3.8) and the values of the Model 1 parameters (Table B.6).

B.7 Derivation of a Dwell time Distribution Function for General f|b Steps

To examine modified initial conditions in our experiments, we considered a simple kinetic model shown in figure B.7. The connection matrix for this model is:

$$M = \begin{bmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -k_2 - k_{-1} & 0 \\ 0 & k_2 & 0 \end{bmatrix}. \tag{B.11}$$

We solve the master equation (eq. B.2) subject to the initial conditions that at time 0 we start in state B:

$$\vec{p}(t=0) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{B.12}$$

After some algebra, the dwell time distribution is shown to be:

$$\frac{dq}{dt}(t) = k_2 p_2(t) = \eta \cdot \lambda_+ e^{-\lambda_+ \cdot t} + (1-\eta) \cdot \lambda_- e^{-\lambda_- \cdot t} \tag{B.13}$$

where $\eta = k_2 \cdot \frac{1+\frac{k_1}{\lambda_+}}{2|D|}$, D^2 is the discriminant of the characteristic equation, $Det(I\lambda - M) = 0$, and $-\lambda_\pm$ are the non-zero eigenvalues of M. By averaging over equation B.13 for Model 1 (Fig. 6A) in the absence of ADP we can show:

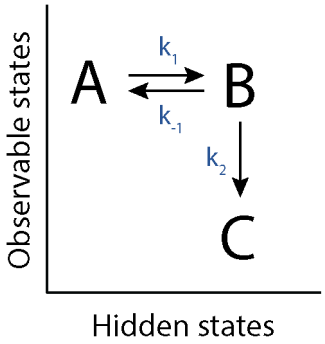

Figure B.7: A kinetic model to analyze f|b steps.

$$\langle t \rangle = \int_0^\infty t \cdot \frac{dq}{dt} \cdot dt = \frac{K_{f|b} + [ATP]}{V_{f|b} \cdot [ATP]}, \tag{B.14}$$

where $V_{f|b} = k_2$ and $K_{f|b} = k_{-1}/k_1$. In the limit that $k_3 \gg k_2$ and $k_{-1} \gg k_2$ we have that $K_{f|f} \approx K_{f|b}$ and $V_{f|f} \approx V_{f|b}$. Evaluating the parameters for this model (discussion s11) suggests that this is the case, which could potentially explain the similarity of the curves in figure 3.6.

B.8 Analysis of f|f and f|b [ATP]-independent Dwell time Distributions Using the AIC

We analyzed the dwell time distributions of [ATP]-independent steps using the corrected Akaike Information criterion (AIC [100]) to analyze f|f and f|b steps. Information (i.e. the predictive power of the model) is lost when data is used to approximate a 'true' underlying distribution function, and when overly complex models are applied to describe the data. The AIC chooses a model by minimizing the information loss of candidate models. The AIC is given by:

$$AIC = \frac{k \cdot (k+1)}{N-k-2} + k - log(L(\hat{\theta}|t)), \tag{B.15}$$

where k is the number of parameters in the model, N is the number of measured data points, L is the likelihood function, and $\hat{\theta}$ is the maximum likelihood estimator for the parameters of the model. The model with the smallest value of the AIC is the one which minimizes the information loss of the data out of the candidate models. We analyzed the following four classes of distribution function:

$$\frac{dp}{dt}(t|a) = a \cdot e^{-a \cdot t} \tag{B.16}$$

$$\frac{dp}{dt}(t|a,b) = \frac{a \cdot b}{a-b} \cdot (e^{-b \cdot t} - e^{-a \cdot t}) \tag{B.17}$$

$$\frac{dp}{dt}(t|a,b,c) = a \cdot b \cdot c \left[\frac{e^{-a \cdot t}}{(c-a)(b-a)} + \frac{e^{-b \cdot t}}{(c-b)(a-b)} + \frac{e^{-c \cdot t}}{(a-c)(b-c)} \right] \tag{B.18}$$

$$\frac{dp}{dt}(t|a,b,c) = a \cdot b \cdot e^{-b \cdot t} + (1-a) \cdot c \cdot e^{-c \cdot t} \tag{B.19}$$

Equations B.16-B.18 are the convolutions of 1, 2 and 3 exponentials, respectively, as would be expected for a Markov chain model [101]. Equation B.19 is a possible model for f|b steps (see section B.7).

Because the dwell time distribution changes along DNA position for f|f and f|b steps (Fig. 5), we calculate the maximum likelihood estimators for each DNA position (i) and each class of distribution function (j), $\hat{\theta}_{ij}$, and sum the AICs for each DNA position together to compute the total information loss, and compare the models. That is:

$$AIC_{total,model\ j} = \Sigma_{i=1}^{N_{steps}} AIC_{ij}. \tag{B.20}$$

The values for the $AIC_{total,modelj}$ are displayed in table B.3. From this table we conclude that, out of the four candidate distribution functions, the information loss is minimized for f|f steps by the convolution of two exponetial distributions, suggesting that there are two rate-limiting steps in the f|f [ATP]-independent step. In contrast, for f|b steps the information loss is minimized by the mixed-exponential model B.19.

Model	k	$AIC_{f\|f}$	$AIC_{f\|b}$
single exponential (eq B.16)	1	-3617.0	-281.4
2 convolved exponentials (eq B.17)	2	-4264.4	-278.3
3 convolved exponentials (eq B.18)	3	-4263.5	-267.3
mixed exponentials (eq B.19)	3	-3694.2	-297.7

Table B.3: Using the AIC to Analyze f|f and f|b [ATP]-independent Steps

The AIC values for the [ATP]-independent f|f and f|b steps for several candidate distribution functions. The value which minimizes the AIC minimizes the information loss by assuming the given model. In total $N_{f|f} = 10052$ and $N_{f|b} = 365$ dwell time measurements of the $f|f$ and $f|b$ steps were used, respectively.

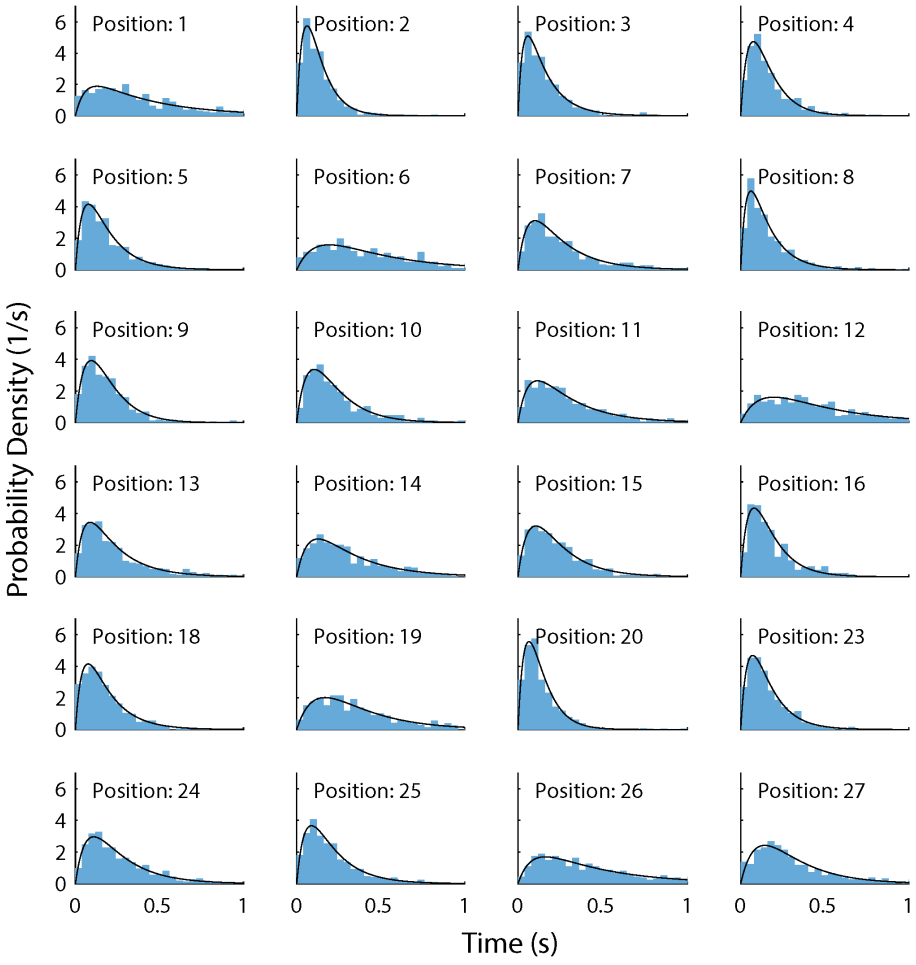

Figure B.8: All f|f [ATP]-independent Dwell Time Distributions

The distribution of dwell times for [ATP]-independent f|f steps (light blue), constructed by taking data from each ATP and ADP titration experiment. The best-fit curve to equation B.17 is plotted on top in black.

B.9 Voltage and Temperature Variation

The force dependence of the dwell time $\langle t \rangle(F)$ in motor enzymes is typically described by the following formula [102]):

$$\langle t \rangle(F) = A \cdot exp(\frac{\Delta E + F \cdot \delta x}{RT}), \tag{B.21}$$

where A is a prefactor with units of time, ΔE is the activation energy of the reaction, F is the force applied to the DNA against the direction of motion of the motor, δx is the characteristic enzyme step size, and RT is the temperature in units of energy. In SPRNT we assume that the electrostatic force on the DNA is proportional to the applied voltage V:

$$\langle t \rangle(V) = A \cdot exp(\frac{\Delta E + \alpha \cdot V}{RT}), \tag{B.22}$$

where α has units of charge. At constant temperature we can write:

$$\langle t \rangle(V) = A' \cdot exp(\beta \cdot V). \tag{B.23}$$

Figures 3.9a and 3.9b show the average dwell time of f|f steps for both [ATP]-dependent and [ATP]-independent steps, averaged over DNA position, plotted against voltage at $[ATP] = 500\ \mu M$ and $[ATP] = 50\ \mu M$, respectively. The results of fitting equation B.23 to the data are displayed in table B.4. In each case we find $\beta \approx 0$, suggesting that the kinetics are independent of voltage in the range of voltages applied.

Setting $\alpha = 0$ in equation B.22 gives:

$$\langle t \rangle = A \cdot exp(\frac{\Delta E}{RT}). \tag{B.24}$$

Figure 3.9c shows the average dwell time of f|f [ATP]-dependent and [ATP]-independent steps versus the inverse temperature of the solution. Fits to equation B.24 yield $\Delta E_{[ATP]-dep} = 60 \pm 11\ kJ \cdot mol^{-1}$ and $\Delta E_{[ATP]-indep} = 77 \pm 15\ kJ \cdot mol^{-1}$. Becasue there are multiple chemical substates within both the [ATP]-dependent and [ATP]-independent pathways, and because we average over DNA position in constructing these curves, these numbers represent

State	$[ATP]$ (μM)	β (mV^{-1})
[ATP]-dep	500	0.001 ± 0.002
[ATP]-indep	500	0.001 ± 0.001
[ATP]-dep	50	-0.002 ± 0.005
[ATP]-indep	50	0.0001 ± 0.004

Table B.4: Best fit value of β to equation B.23 for the curves shown in figure 3.9.

approximately the average activation energy of the rate-limiting step for each observable step type.

B.10 Calculation of Average Dwell Time of f|f [ATP]-dependent Steps Using the Steady-state Approximation

The goal of this section is to calculate the average dwell time of f|f [ATP]-dependent steps for both Model 1 and Model 2 as a function of [ATP] and [ADP]. Because there are four states in Model 1, the eigenvalues of the connection matrix come from a cubic polynomial, and solutions are difficult to calculate. Similarly, for Model 2 there are five states, which requires solving a quartic polynomial to obtain the eigenvalues. Thus, we seek to rewrite this problem to make use of the steady-state approximation [41], which can be used to determine average dwell times by solving a linear system of equations. First we note that none of the processes which determine the dwell time of the [ATP]-independent step can affect the average dwell time of the [ATP]-dependent step, so we compress the rate constants $k_{\pm H}$, $k_{\pm P}$, and k_1 into a single rate parameter Ω. Restricting ourselves to f|f steps means we can ignore k_{-2}. We use the fact that for the [ATP]-dependent step $p_{b|f} \ll 1$ to conclude that $k_{-1} \ll k_D$ (Model 1) or $k_{-1} \ll k_T[ATP]$ (Model 2), so we can ignore k_{-1} as well. Figure B.9 summarizes each of these observations into a single kinetic path for Model 1. In this form, we can use the steady-state approximation so that:

$$M \cdot \vec{p}_{ave} = \vec{0}, \tag{B.25}$$

where M is the connection and $\vec{p}_{ave}$ is the average probability that a given state is occupied. The normalization condition is:

$$\Sigma_i \; p_{ave,i} = 1, \tag{B.26}$$

where i indexes the entries of $\vec{p}_{ave}$. The connection matrix for Model 1 in this approximation is:

$$M_{Model\ 1} = \begin{bmatrix} -k_D & k_{-D}[ADP] & 0 & \Omega \\ k_D & -k_{-D}[ADP] - k_T[ATP] & k_{-T} & 0 \\ 0 & k_T[ATP] & -k_{-T} - k_2 & 0 \\ 0 & 0 & k_2 & -\Omega \end{bmatrix} \tag{B.27}$$

Equations B.26 and B.27 are used to solve for $p_{ave,i}$. Following Keller[41], the total reaction rate for Model 1 is the average probability that the ATP bound state of Hel308 is occupied multiplied by the transition rate:

$$r_{total,Model\ 1} = \frac{1}{\langle t \rangle_{total,Model\ 1}} = k_2 \cdot p_{ave,Hel308 \cdot ATP}, \tag{B.28}$$

We are interested in the rate of just the $f|f$ [ATP]-dependent step. The average total time to progress through the entire pathway is the sum of the time to go through the [ATP]-dependent step plus the time to go through the [ATP]-independent step:

$$\langle t \rangle_{total,Model\ 1} = \langle t \rangle_{[ATP]_d \to [ATP]_i} + \langle t \rangle_{[ATP]_i \to [ATP]_d} = \langle t \rangle_{[ATP]_d \to [ATP]_i} + \frac{1}{\Omega} \tag{B.29}$$

The transition rate for [ATP]-dependent steps can be solved by rearranging equation B.29.

$$rate_{[ATP]_d \to [ATP]_i} = \frac{1}{\langle t \rangle_{[ATP]_d \to [ATP]_i}} = (\langle t \rangle_{total} - \frac{1}{\Omega})^{-1} \tag{B.30}$$

Plugging the results of B.26, B.27 and B.28 into B.30 for Model 1 yields:

$$rate_{Model\ 1} = \frac{V \cdot [ATP]}{K + [ATP] + d \cdot [ADP]} \tag{B.31}$$

Applying an identical calculation to Model 2 gives:

$$rate_{Model\ 2} = \frac{V' \cdot [ATP]}{K' + [ATP] + d' \cdot [ADP] + e \cdot [ATP] \cdot [ADP]} \tag{B.32}$$

where $K = \frac{k_{-T}+k_2}{k_T} \cdot \frac{k_D}{k_2+k_D}$, $V = \frac{k_2 \cdot k_D}{k_2+k_D}$, $d = K \cdot \frac{k_{-D}}{k_D}$, $V' = \frac{k_T k_2 k_D k_H}{D}$, $K' = \frac{k_{-T}k_2k_D + k_{-T}k_2k_{-H} + k_2k_Dk_H}{D}$, $d' = \frac{k_{-D}k_{-T}k_{-H}}{D}$, $e = \frac{k_{-D}k_T(k_H+k_{-H})}{D}$, with $D = k_T k_2 k_D + k_T k_2 k_{-H} +$

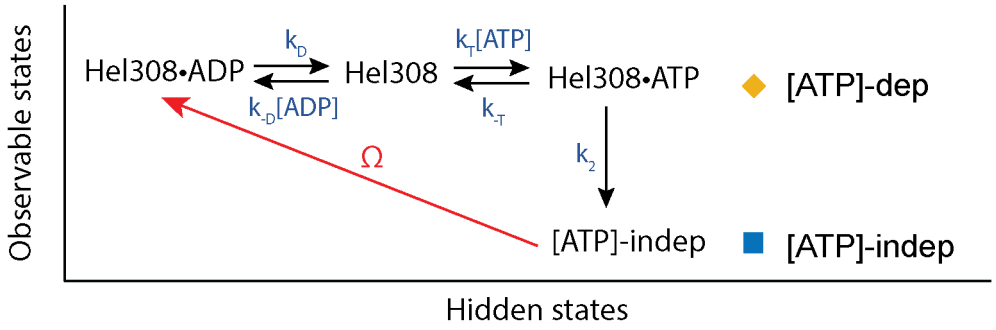

Figure B.9: The kinetic model from main text figure 6a, reduced to a form that allows easy application of the steady-state approximation.

$k_T k_2 k_H + k_T k_D k_H$. Equations B.31 and B.32 are independent of Ω, as they must be. These expressions differ qualitatively only by the existence of the term $e \cdot [ATP] \cdot [ADP]$ that couples the ATP and ADP concentrations in equation B.32.

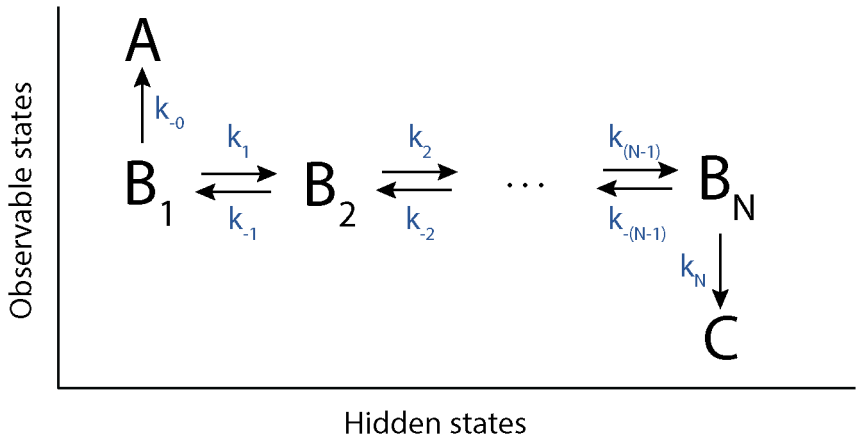

Figure B.10: Calculating the Probability of b|f [ATP]-dependent Steps as a function of [ATP] and [ADP]

A hypothetical model to illustrate how to calculate the backstep probability in a chain Markov model.

B.11 Derivation of the Probability of a b|f Step for [ATP]-dependent Steps in Model 1 and Model 2

Because SPRNT gives us access to both forwards and backwards steps, we sought to derive a general formula for the probability of a backstep in terms of the underlying rate constants of a given kinetic model. Consider the hypothetical kinetic model shown in figure B.10. Assume that at time 0 the enzyme is in the state B_1. In the first passage time problem, the only states that can be occupied at time ∞ are A and C, corresponding to a backwards step and a forwards step, respectively. The probability of a backstep is thus:

$$p_{back} = \lim_{t\to\infty} p_A(t). \tag{B.33}$$

As done previously, we solve the equation $\frac{d\vec{p}}{dt}(t) = M \cdot \vec{p}(t)$, however in this case we use the Laplace transform method. We define $\vec{p}(s) = \mathcal{L}(\vec{p}(t))$ to be the laplace transform of $\vec{p}(t)$. The solution in transform space is easily shown to be:

$$\vec{p}(s) = (sI - M)^{-1} \cdot \vec{p}(t = 0), \tag{B.34}$$

where I is the identity matrix. The inverse transformation is done by the Bromwich integral [103]:

$$p_j(t) = \Sigma_{i=1}^{N_{poles}} Res_{s=s_i}[p_j(s) \cdot e^{s \cdot t}] \tag{B.35}$$

where j indexes the entries of $\vec{p}$. $Res_{s=s_i}[p_j(s)e^{s \cdot t}]$ is the residue evaluated at the poles s_i of $p_j(s)$, and s is taken as a complex variable. The backstep probability can be written using equations B.33 and B.35 as:

$$p_{back} = \lim_{t \to \infty} \Sigma_{i=1}^{N_{poles}} Res_{s=s_i}[p_A(s) \cdot e^{s \cdot t}]. \tag{B.36}$$

This sum is over a finite number of poles, so we interchange the limit and residue expressions, and because $p_A(s)$ is not a function of time:

$$p_{back} = \Sigma_{i=1}^{N_{poles}} Res_{s=s_i}[p_A(s) \cdot \lim_{t \to \infty} e^{s_i \cdot t}]. \tag{B.37}$$

Note that we have explicitly evaluated the residue in the exponential term $e^{s_i \cdot t}$. Next we use the fact that the poles of $p_i(s)$ are eigenvalues of the matrix M, which in the first-passage time problem have the property that $s_i \leq 0$. The limit as $t \to \infty$ vanishes for all negative eigenvalues, collapsing the sum and leaving us with the simple expression:

$$p_{back} = Res_{s=0}[p_A(s)], \tag{B.38}$$

where $P_A(s)$ is determined from equation B.34. If the pole at $s = 0$ is first order then this expression reduces to:

$$p_{back} = \lim_{s \to 0} s \cdot p_A(s). \tag{B.39}$$

Other expressions have been derived for the backstep probability [83, 84], but to our knowledge, the form of equation B.39 has not been derived. This expression is simple to apply, because both the matrix inversion and residues of rational functions are easy to evaluate. Applying equation B.39 to Model 1 and Model 2 yields:

$$p_{back,Model\ 1} = \frac{k_{-1}k_{-D}(k_2 + k_{-T})[ADP] + k_{-1}k_2k_T[ATP]}{k_{-1}k_{-D}(k_2 + k_{-T})[ADP] + (k_{-1} + k_D)k_2k_T[ATP]}, \tag{B.40}$$

$$p_{back,Model\ 2} = \frac{k_{-1}(k_2k_Dk_H + k_{-T}k_2k_D + k_2k_{-H}k_{-T} + k_{-D}k_{-H}k_{-T}[ADP])}{k_{-1}(k_2k_Dk_H + k_{-T}k_2k_D + k_2k_{-H}k_{-T} + k_{-D}k_{-H}k_{-T}[ADP]) + k_2k_Dk_Hk_T[ATP]}. \tag{B.41}$$

To apply equation B.39 to Model 1 with the additional diffusion term, we simply need to note that in this model we have $p_{back} = \lim_{t\to\infty}(p_{Hel308^*\cdot ADP} + p_{Hel308^*})$, giving:

$$p_{back,Model\ 1+diffusion} = \frac{2k_{dif}(k_{-1} + k_D)(k_2 + k_{-T}) + k_{-1}k_{-D}(k_2 + k_{-T})[ADP] + k_{-1}k_2k_T[ATP]}{2k_{dif}(k_{-1} + k_D)(k_2 + k_{-T}) + k_{-1}k_{-D}(k_2 + k_{-T})[ADP] + (k_{-1} + k_D)k_2k_T[ATP]}. \tag{B.42}$$

First, we note that in the limiting case of $[ADP] \to 0$ in equation B.40, $p_{back,Model\ 1} \to \frac{k_{-1}}{k_{-1}+k_D}$, as must be the case for a simple branched pathway. In order to do fitting with these models, we rearrange expressions B.40-B.42 into more accessible forms using the forms of V,K,d,V',K' and d' from equations B.31-B.32:

$$p_{back,Model\ 1} = \frac{d[ADP] + \frac{V}{k_D}[ATP]}{d[ADP] + \frac{V}{k_D}[ATP](1 + \frac{k_D}{k_{-1}})} \tag{B.43}$$

$$p_{back,Model\ 2} = \frac{K' + d'[ADP]}{K' + d'[ADP] + \frac{V'}{k_{-1}}[ATP]} \tag{B.44}$$

$$p_{back,Model\ 1+diffusion} = \frac{2\frac{k_{dif}}{k_D}(1 + \frac{k_D}{k_{-1}})K + d[ADP] + \frac{V}{k_D}[ATP]}{2\frac{k_{dif}}{k_D}(1 + \frac{k_D}{k_{-1}})K + d[ADP] + \frac{V}{k_D}[ATP](1 + \frac{k_D}{k_{-1}})}. \tag{B.45}$$

We have managed to write B.40-B.42 in terms of the same proportionality constant (d or d') with which the average dwell time depends on $[ADP]$ at fixed $[ATP]$. It is important to specify at least one of these parameters when doing fits to equations B.43-B.45, because

without a parameter to set the scale, there will always be some degeneracy, due to the fact that p_{back} will be unchanged by scaling the concentrations and rate constants by a constant factor. Ultimately, for model selection, we are interested in the quality of the fit given a certain value of d, so we use $d = 0.003$, obtained from averaging over each DNA position in table B.5. Fits to equations B.43-B.45 to the data p_{back} vs. $[ATP]$ and $[ADP]$ for all experiments are displayed in figure B.12. Of the three models considered here, Model 1 together with the diffusion term (black line) best fits the data.

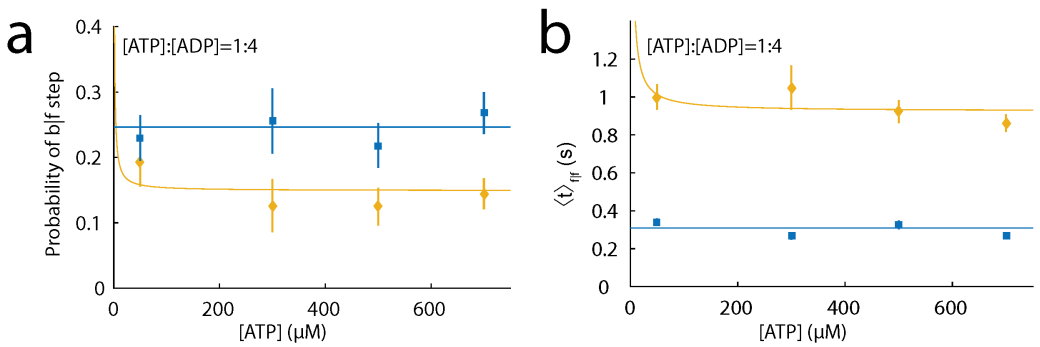

Figure B.11: Fixed Ratio [ATP]:[ADP] = 1:4 Experiment

(a) Probability of a backstep for the [ATP]-dependent (yellow) and [ATP]-independent (blue) steps averaged over DNA position vs. [ATP] at fixed ratio [ATP]:[ADP] = 1:4. Weighted averages to the data are plotted on top. (b) Average dwell time of f|f [ATP]-dependent (yellow) and [ATP]-independent (blue) step averaged over DNA position vs. [ATP] at fixed ratio [ATP]:[ADP] = 1:4. Best fit to main text equation 3 is plotted on the [ATP]-dependent step, yielding $e^* \approx 0$, in line with Model 1. The weighted average (blue) is plotted on the [ATP]-independent step.

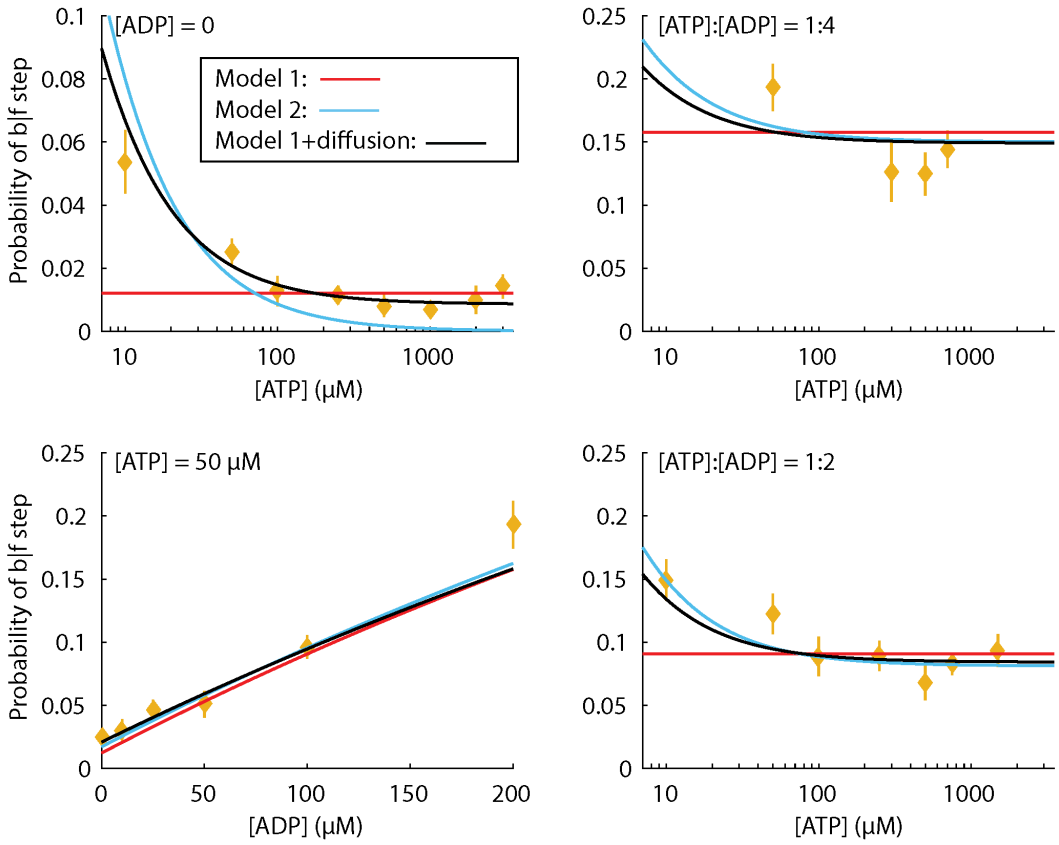

Figure B.12: Comparing Kinetic Models by Analyzing Probability of b|f Steps

Probability of a $b|f$ step in several different experiments: (top left) $[ATP]$ varied, $[ADP] = 0$. (bottom left) $[ATP] = 50\mu M$, $[ADP]$ varied. (top right) $[ATP]$ and $[ADP]$ varied at fixed $[ATP] : [ADP] = 1{:}4$. (bottom right) $[ATP]$ and $[ADP]$ varied at fixed $[ATP] : [ADP] = 1{:}2$. Red, blue and black lines are simultaneous fits to the data in each of the panels shown for Model 1 (red, eq. B.43, $\chi^2/\nu = 3.0$), Model 2 (blue, eq. B.44, $\chi^2/\nu = 2.9$) and Model 1 with diffusion (black, eq. B.45, $\chi^2/\nu = 1.0$), respectively.

B.12 Estimation of Kinetic Parameters for [ATP]-dependent Steps

To gain further insight into Hel308 kinetics, we calculated the relevant kinetic parameters for the [ATP]-dependent step using likelihood maximization. From Model 1 (Fig 6a) and the ATP and ADP titration experiments we estimated the 5 parameters that determine the progression of [ATP]-dependent f|f steps ($k_{\pm T}, k_2, k_{\pm D}$) using maximum likelihood methods. Here we assume that $k_{-1} \ll k_D$, an assumption that is justified by the low probability of a backstep for [ATP]-dependent steps in the absence of ADP ($p_{back} < 0.01$ for most DNA positions, so any errors made under this assumption are small compared to the errors on $V_{f|f}$ and $K_{f|f}$). The probability distribution of dwell times for a kinetic model is determined by numerically solving the master equation with connection matrix:

$$M = \begin{bmatrix} -k_D & k_{-D} \cdot [ADP] & 0 & 0 \\ k_D & -k_{-D} \cdot [ADP] - k_T \cdot [ATP] & k_{-T} & 0 \\ 0 & k_T \cdot [ATP] & -k_2 - k_{-T} & 0 \\ 0 & 0 & k_2 & 0 \end{bmatrix}. \tag{B.46}$$

For f|f steps the master equation is subject to the initial conditions:

$$\vec{p}(t=0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{B.47}$$

The observable dwell time distribution, $\frac{dq}{dt}$, is given by:

$$\frac{dq}{dt}(t|k_D, k_{-D}, k_T, k_{-T}, k_2) = k_2 \cdot p_3(t). \tag{B.48}$$

As is, the model has five free parameters. These can be reduced to two parameters by using the measured values of $K_{f|f}$, $V_{f|f}$, and K_I, defined by the expression $K_I \equiv K_{f|f}/d$. Using the results of section B.10 we showed:

$$K_{f|f} = \frac{k_2 + k_{-T}}{k_T} \cdot \frac{k_D}{k_D + k_2}, \tag{B.49}$$

$$V_{f|f} = k_2 \cdot \frac{k_D}{k_D + k_2}, \tag{B.50}$$

$$K_I = \frac{k_D}{k_{-D}}. \tag{B.51}$$

From these expressions we solve for k_T, k_D and k_{-D} in terms of k_2, k_{-T} and the measured parameters:

$$k_D = V_{f|f} \cdot \frac{k_2}{k_2 - V_{f|f}}, \tag{B.52}$$

$$k_T = \frac{k_{-T} + k_2}{K_{f|f}} \cdot \frac{V_{f|f}}{k_2}, \tag{B.53}$$

$$k_{-D} = \frac{k_D}{K_I}. \tag{B.54}$$

Using these expressions, the matrix M depends only on k_2 and k_{-T}. To estimate these parameters, we evaluate the log likelihood function on a two-dimensional grid spanned by guess values of k_2 and k_{-T}:

$$log(L(k_2, k_{-T}|t)) = \Sigma_{i=1}^{N} \Sigma_{j=1}^{n_i} log(\frac{dq}{dt}(t_j|k_2, k_{-T})), \tag{B.55}$$

where i indexes the sum over each experimental condition and j indexes a sum over each measured data point at those conditions. We only use the ATP titration experiments at [ADP] = 0 in the likelihood analysis and then use equation B.54 to solve for k_{-D}. Figure B.13 shows the result of this calculation for several DNA positions. For most DNA positions there is a clear likelihood peak around the values of k_2 and k_{-T} which maximize the likelihood function.

To estimate the errors we note that both $K_{f|f}$ and $V_{f|f}$ have measurement errors associated with their values. We repeated the calculation B.55 for the log likelihood 200 separate times for each DNA position by monte carlo sampling the joint distribution of $K_{f|f}$ and $V_{f|f}$

to build the distribution of possible values of $k_{\pm T}, k_2$ and $k_{\pm D}$. For each monte carlo sample we extracted the values of k_2 and k_{-T}. For each sample of k_2 and k_{-T} we used equations B.52-B.54 to calculate the other model parameters (Table B.6). The collection of monte carlo samples are distributions of the model parameter values. We report the mean and standard deviation of these distributions in table B.5. The log likelihood at several DNA positions (6.5,13.5,23.5) did not decay at increasing k_{-T}. This may be because the distribution of dwell times (equation B.55) is sensitive to k_{-T} at low $[ATP]$ [85], however much of the data was obtained at higher concentrations. More data at low $[ATP]$ would likely lead to a better resolved peak.

We examined a variant of this model, in which ATP directly induces a transition from the [ATP]-dependent step to the [ATP]-independent step (figure B.14). In the absence of ADP we can write the dwell time distribution for f|f steps as:

$$\frac{dq}{dt}(t) = \frac{V \cdot [ATP]}{[ATP] - K}(e^{-V \cdot t} - e^{-\frac{V \cdot [ATP] \cdot t}{K}}). \tag{B.56}$$

This model has no free parameters. Repeating the calculation of equation B.55 with the dwell time distribution B.56, and evaluating the AIC for each model, we find that for Model 1 the AIC is -8868 and for the alternative model the AIC is -8671, suggesting that Model 1 better describes the data.

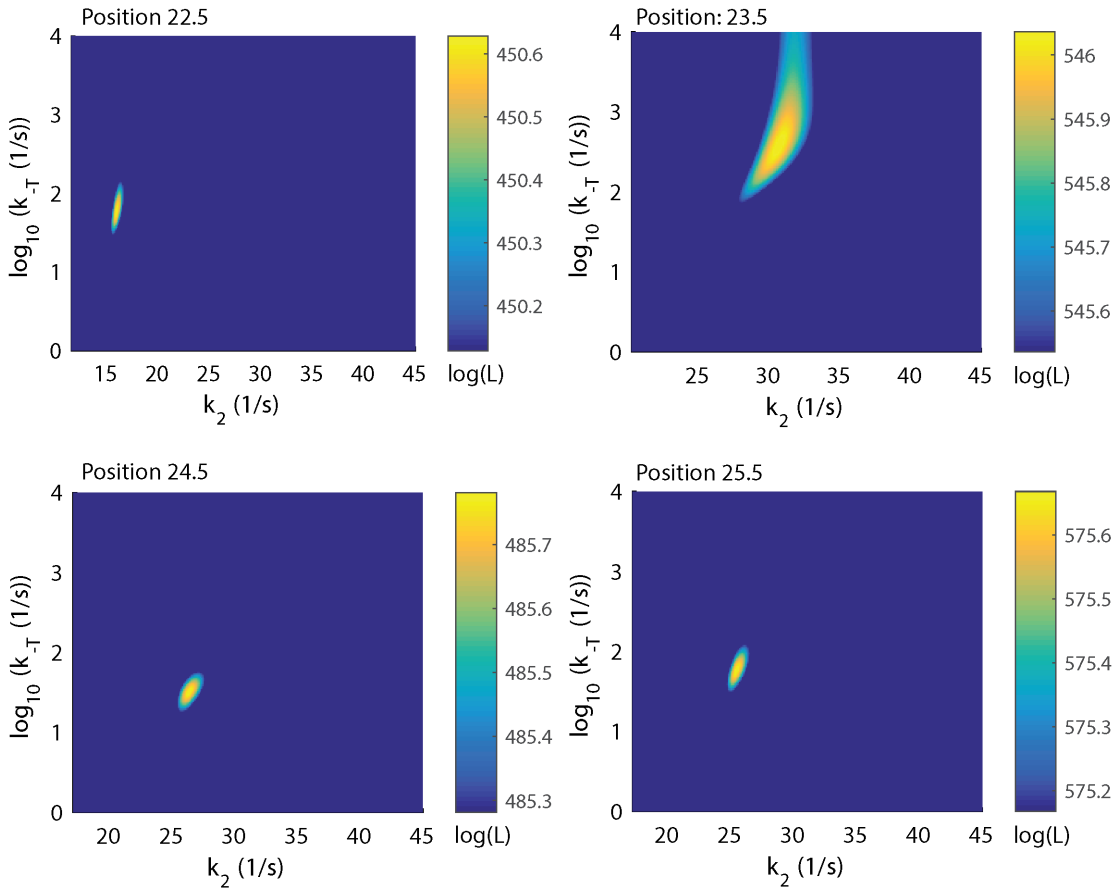

Figure B.13: Calculating Kinetic Parameters

Log likelihood function (equation B.55) for several [ATP]-dependent steps to determine the values of the parameters k_2 and k_{-T}. The x-axis shows test values of k_2, the y-axis is the log of test values of k_{-T}, and the color axis is the log Likelihood for a single realization of the Monte-Carlo values of $K_{f|f}$ and $V_{f|f}$. The peak in yellow is evidence of a single maximum value. The log likelihood function at DNA position 23.5 did not decay with increasing k_{-T}, possibly indicating a lack of data at low $[ATP]$.

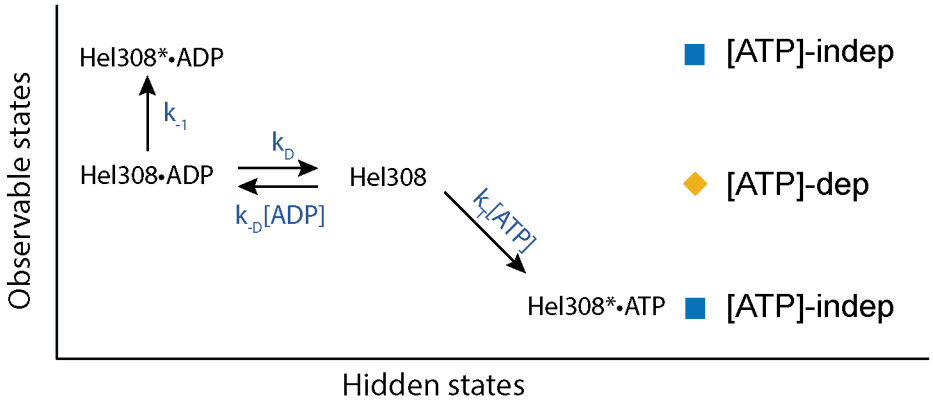

Figure B.14: Alternative Model 1

An alternative Model 1. Rather than binding as a hidden step, ATP directly induces the conformational change of Hel308.

DNA position (nt)	$K_{f\|f}(\mu M)$	$V_{f\|f}(s^{-1})$	$K_I(\mu M)$
0.5	82 ± 24	11.3 ± 1.5	31 ± 14
1.5	92 ± 44	10.1 ± 2.1	68 ± 48
2.5	412 ± 125	23.5 ± 3.6	81 ± 36
3.5	329 ± 40	11.0 ± 0.7	62 ± 19
4.5	181 ± 41	28.6 ± 3.2	47 ± 17
5.5	174 ± 44	12.8 ± 1.5	30 ± 9
6.5	153 ± 35	13.9 ± 1.2	30 ± 9
7.5	237 ± 67	12.4 ± 1.6	40 ± 17
8.5	100 ± 33	16.0 ± 2.3	18 ± 8
9.5	214 ± 59	23.6 ± 3.8	57 ± 25
10.5	241 ± 51	19.0 ± 1.9	65 ± 24
11.5	143 ± 55	21.0 ± 3.8	46 ± 23
12.5	110 ± 21	16.4 ± 1.3	37 ± 16
13.5	228 ± 82	22.0 ± 3.6	37 ± 17
14.5	66 ± 24	10.2 ± 1.5	49 ± 25
15.5	117 ± 51	23.4 ± 3.8	35 ± 22
17.5	75 ± 12	9.6 ± 0.7	21 ± 6
18.5	215 ± 40	34.0 ± 3.0	57 ± 13
19.5	108 ± 27	15.0 ± 1.6	37 ± 12
22.5	97 ± 18	14.7 ± 1.1	42 ± 16
23.5	250 ± 46	25.4 ± 2.5	55 ± 24
24.5	255 ± 62	23.3 ± 3.0	73 ± 26
25.5	140 ± 70	22.9 ± 4.4	40 ± 27
26.5	152 ± 26	26.4 ± 1.8	33 ± 12

Table B.5: List of Michaelis-Menten Parameters

[ATP]-dependent step kinetic parameters as defined in section B.10.

DNA position (nt)	$k_T(\mu M^{-1} \cdot s^{-1})$	$k_{-T}(s^{-1})$	$k_2(s^{-1})$	$k_D(s^{-1})$	$k_{-D}(\mu M^{-1} \cdot s^{-1})$
0.5	0.36 ± 0.04	19 ± 3	11.6 ± 1.2	179 ± 11	5.7 ± 2.4
1.5	0.56 ± 0.22	34 ± 10	9.9 ± 2.0	247 ± 54	3.8 ± 2.4
2.5	0.28 ± 0.07	99 ± 3	25.9 ± 3.4	303 ± 28	3.7 ± 1.7
3.5	0.14 ± 0.03	36 ± 3	12.2 ± 1.5	114 ± 14	1.8 ± 0.5
4.5	0.61 ± 0.08	85 ± 8	30.3 ± 2.9	410 ± 24	9.1 ± 3.4
5.5	0.28 ± 0.04	38.8 ± 2.4	13.9 ± 1.4	162 ± 12	5.3 ± 1.6
6.5*	4.4 ± 1.5	709 ± 94	15.5 ± 1.5	127 ± 5	4.3 ± 1.2
7.5	0.59 ± 0.18	125 ± 10	12.8 ± 1.9	226 ± 16	5.7 ± 2.4
8.5	0.26 ± 0.02	12 ± 2	19.1 ± 1.3	103 ± 5	5.8 ± 2.4
9.5	0.44 ± 0.03	76 ± 9	24.8 ± 2.9	323 ± 11	5.6 ± 2.4
10.5	0.27 ± 0.04	57 ± 3	22.8 ± 2.6	118 ± 5	1.8 ± 0.7
11.5	0.28 ± 0.03	22 ± 2	23.7 ± 2.1	167 ± 8	3.6 ± 1.8
12.5	0.40 ± 0.06	30 ± 2	18.4 ± 1.6	140 ± 7	3.8 ± 1.6
13.5*	0.87 ± 0.28	184 ± 33	23.9 ± 2.9	265 ± 12	7.0 ± 3.1
14.5	0.39 ± 0.06	17 ± 2	11.1 ± 1.0	119 ± 9	2.4 ± 1.2
15.5	0.30 ± 0.05	12 ± 2	26.2 ± 2.0	274 ± 30	7.8 ± 4.8
17.5	0.29 ± 0.03	13 ± 2	10.6 ± 1.2	103 ± 5	5.0 ± 1.5
18.5	0.30 ± 0.04	35 ± 2	38.7 ± 3.3	240 ± 27	4.2 ± 0.8
19.5	0.38 ± 0.07	30 ± 2	18.0 ± 1.5	118 ± 5	3.2 ± 1.0
22.5	0.69 ± 0.11	58 ± 4	16.2 ± 1.4	172 ± 8	4.1 ± 1.5
23.5*	1.26 ± 0.12	346 ± 45	30.6 ± 2.9	189 ± 8	3.5 ± 1.6
24.5	0.21 ± 0.02	30 ± 3	24.9 ± 2.7	228 ± 9	3.1 ± 1.1
25.5	0.56 ± 0.08	56 ± 6	24.0 ± 2.2	370 ± 20	9.3 ± 6.2
26.5	0.46 ± 0.06	51 ± 4	31.0 ± 2.3	181 ± 9	5.4 ± 1.9

Table B.6: Calculated Parameters For Model 1

[ATP]-dependent step kinetic parameters as in figure B.9, calculated from equation B.55. The likelihood function for those steps with an asterisk next to them did not decay as $k_{-T} \to \infty$, suggesting that the values of k_{-T} and k_T cannot be trusted.

Appendix C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

C.1 Materials and Methods

The experimental conditions and analysis are identical to what was done in chapter 3, with $[ATP] = 1000\ \mu M$, and $T = 37\ ^{\circ}C$. The DNA sequence used was:

5′ PTACTACTACATTACXXCTTTGTCGTTGTGCAGTCGTT...
...NNNNNNNNNNNNNNNNNTGGTATCTCACTATCGCATTCTCATGCAGGTCGTAGCC 3′

with complement sequence:

5′ CCTGCATGAGAATGCGATAGTGAGAYYYYZ. 3′

C.2 Calculation of p-value using 2-sample KS-Test

We sought to quantify the relative probability that the observed kinetics were due to effects of DNA in the nanopore, or whether they only depended on the position of Hel308 along the DNA. We used the 2-sample Kolmogorov-Smirinov test (KS test) to assign a p-value that a given pair of empirical histograms are sufficiently different given that they were drawn from the same underlying distribution function, regardless of the 'true' underlying PDF. We calculate the p-value for each pair of histograms in figure 4.1c (including those [ATP]-dependent steps in between the displayed [ATP]-independent steps), and calculate the total log probability by:

$$log(P) = \Sigma_i log(p_i). \tag{C.1}$$

The relative probabilities that the histograms are drawn from the same underlying distribution when aligned based on ion current as opposed to distance along the DNA to be 10^{-11},

N_{21} Sequence ($5' \rightarrow 3'$)
AAAAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTTTT
CCCCCCCCCCCCCCCCCCCC
CCTCAAAATCAGATCTCACTA
CCTCAAAATCACATCTCACTA
CCTCAAAATCCCATCTCACTA
CCTCAAAATCCCCTCTCACTA
AGAGAGAGAGACCCCCCCCCC
AGAGAGAGAGAGAGAGAGAGA
CCCCCCCCCCCACCCCCCCCC

Table C.1: List of DNA sequences used in chapter 4

suggesting that the observed dynamics are indeed caused by the DNA sequence in Hel308, as opposed to effects of the nucleotides in the constriction of MspA.